Probing OleA Mechanism and Diversity Using Alternative Substrates

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MINNESOTA BY

Megan Danielle Smith

IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Dr. Lawrence P. Wackett, Advisor

AUGUST, 2021

Acknowledgements

Thank you to my advisor, Dr. Lawrence Wackett, for allowing me to join his lab and giving me a chance to prove myself as a researcher. Thank you. I've learned so much and gained so much confidence in myself and my research. I owe a lot of the growing I have done to my time in your lab and I extend my deepest gratitude for that.

I want to thank everyone I have met in the Wackett lab as well. Everyone in the lab has helped me in both big and small ways. Thank you to my rotation mentors, Dr. James Christenson and Dr. Serina Robinson, who encouraged me when I first joined and whose work is an inspiration and foundation for all that I have accomplished on the project. Thank you to Kelly Aukema for always looking out for me. You answered every question, no matter how frequent and were always someone I could talk to about any issue. I would not have made it as far as I have without you. Thank you to all the undergraduate mentees. By teaching others, we can deepen our own learning, and this has never been truer than when working with some of the brilliant undergraduates I have had the pleasure of mentoring.

Thanks to the friends I have made in my MICaB cohort who I could count on for support and comradery. Thanks to Katrina Jackson, Alex Villarreal, Jordan Neumann, Ruth Lee especially for all the many fun adventures that reminded me that there is more to graduate school than classwork and lab. You all are such smart and accomplished people that I look up to and am so happy to have met. Thank you as well to my friends in other states who supported me as well.

Thank you to my entire family. Thank you to my mom and dad for supporting me through every step, for being there when I struggled with setbacks and for celebrating with me in my successes. Thank you to my family in Minnesota for providing a place to retreat to when I needed a break and for the memories I made while there. Every fun conversation, home-cooked meal, jar of homemade pickles, and fresh garden grown vegetable helped keep me grounded in times of high stress. Having the opportunity to make memories with my grandmother in the home I've known since childhood will mean more than words

can express. I miss you so much. Thank you to my family in Texas who supported from afar with calls and texts. I looked forward to every holiday that I could come down and visit knowing it would always be a good time and leave me refreshed to continue my graduate work. Thanks to my cousins who are always available to talk to when I needed a break from graduate school. I want to especially thank my cousin Dante who spent hours talking to me about every topic imaginable during my first years here. It meant more than I said at the time and saw me through a lonely first year while I adjusted to life in a new city and state. I miss you every day. You may no longer be here to see me finish my degree, but I know you would have been just as excited for me as I am about what I've been able to accomplish.

I could not be here without the support of so many people, many more than I've been able to mention here. It is due to their unwavering support that I am at the place I am today. Thank you all so much.

Dedication

I would like to dedicate this work to my family and friends who have believed in me and supported me through every step of this journey.

Abstract

Various microorganisms posses the ability to create long chain olefinic hydrocarbons, which can be useful as biofuel precursors or for the creation of specialty chemicals. Additionally, the biosynthesis pathways for the production of these olefins can be used to create other high value compounds such as β-lactone therapeutics and surfactants. Our lab has identified hundreds of possible olefin and β-lactone biosynthetic gene clusters, however many of the end products are unknown. In order to test the diversity of olefins produced, we focused our efforts on the first enzyme in the pathway, OleA. OleA is a thiolase that condenses two fatty acyl-CoAs into a β-keto acid, which eventually becomes the olefin or β-lactone end-product. Earlier research has shown that OleA is responsible for the overall shape of the end-product, through its substrate specificity. Thus, we surmised that studying OleA would provide insight on the natural variance of Olefin and β-lactone compounds produced in nature. Through a collaboration with the Joint Genome Institute we have sampled a total of 72 OleA recombinant proteins using a new colorimetric assay that was developed through the use of inexpensive and readily available *p*-nitrophenyl compounds as substrate analogs. Our research has shown that many of these OleA proteins are able to accept a wide variety of compounds, many of which are not found in nature. In addition to the development of the screen, we discovered that the *p*-nitrophenyl ester could only participate as the first substrate of the Claisen condensation reaction of OleA. We then exploited this to produce the β-keto acid product using *p*-nitrophenyl esters, as well as further refine the OleA mechanism, by demonstrating the directionality of the Claisen condensation. Finally, we showed the addition of cetyltrimethylammonium bromide (CTAB), below critical micelle concentrations, can improve the yield of OleA, both when using native or alternative substrates. This improves the potential for OleA to become a biotechnologically important enzyme, and points to future ways we can improve the overall enzyme pathway to generate large amounts of desired product. This

thesis represents a first step in a larger goal of producing industrially relevant β-lactone compounds and their derivatives.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 – Introduction

"It is good to have an end to journey toward; but it is the journey that matters, in the end." – Ursula K. Le Guin (The Left Hand of Darkness)

The Utility of β-lactones in Industry

β-Lactones as Therapeutics

The β-lactone is a four-membered ring made up of a cyclized ester (Figure 1-1). These compounds have high ring strain leading to both a high level of instability and reactivity (1–4). While this instability can make the β-lactone difficult to isolate, structures have been isolated from all domains of life. Currently around 30 natural product core scaffolds containing β-lactones have been described (3, 4), showing anticancer, antitumor, and antiobesity properties (5–10). However, little research has been done with respect to the native function of these products.

One reason for the wide range of uses for β-lactone natural products is that β-lactone rings can be attacked at both the C2 and C4 position (Figure 1-1) (1–4, 11). The ring can be opened at these two sites by cysteine, threonine, or serine residues, leading to covalent inhibition in some cases (3, 4). These features allow β-lactone containing structures to interact with a broad range of enzymes that fall under four enzyme classes: hydrolases, transferases, ligases, and oxidoreductases. Also, β-lactones have been shown in a chemical profiling study to target over 20 different enzymes in these categories (10).

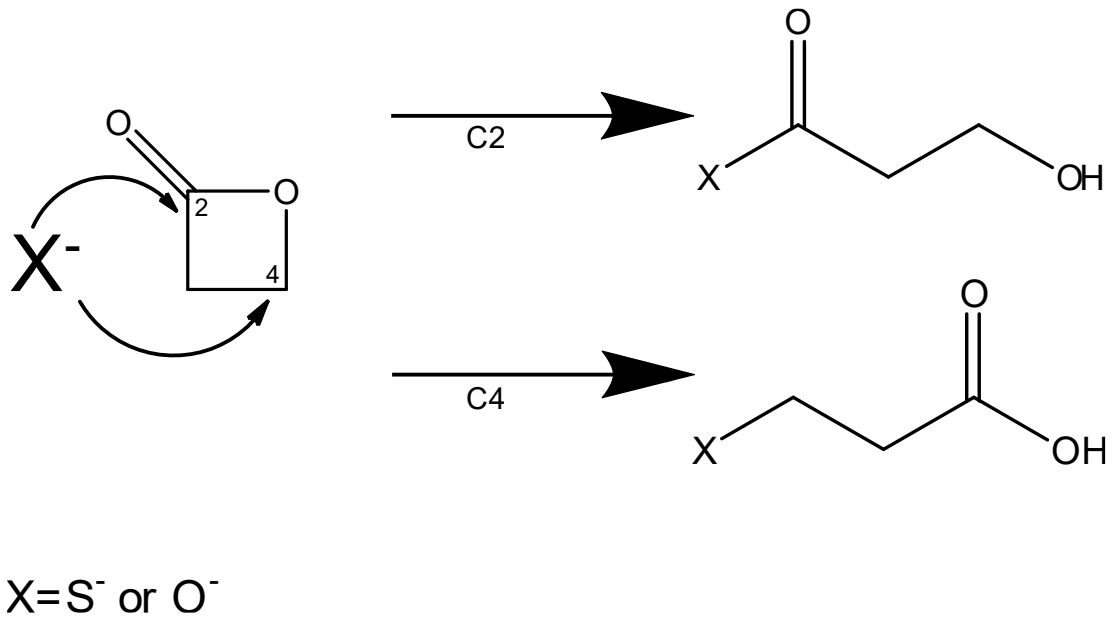


Figure 1-1. Schematic showing β-lactone ring opening at C2 and C4 position

The multitude of enzymatic targets for the β-lactone has made it difficult for these natural products to be developed into drugs, as there exists a possibility for deleterious effects through non-specific interaction with unwanted cellular targets. At the time of this thesis, there is only one FDA approved drug that uses a β-lactone moiety as its mode of action: the antiobesity medication lipstatin (5, 12). This is in contrast to the structurally similar β-lactam which has wide use in medicine today(13, 14). However, the potential for β-lactones as medically relevant compounds still exists. Clinical trials are being held for multiple β-lactone containing therapeutics including salinosporamide A which is in phase III clinical trials for the treatment of multiple-myeloma, a cancer of plasma cells (7).

β-Lactones as Intermediates

Not only do β-lactone compounds have clinical potential, but they have been used as intermediates both in nature and synthetically to create products important to both the host organism and industry, respectively. The highly

reactive strained ring allows for synthesis of a wide range of other bioactive molecules.

There are few known instances of β-lactones used as an intermediate in nature, a place for potential future research. One example of this are membrane olefinic hydrocarbons, which are industrially relevant as a potential fuel source (6, 15). Our lab and others have done extensive research on this biosynthetic pathway in *Xanthomonas campestris,* as well as a few other organisms and is what led to the discovery of the β-lactone synthetase (6, 15–17). Another example is the formation of both polymers and monomers of a vibralactone derivative called vibralactoxime. These represent the first oxime esters discovered in nature (18). These oximes have been shown to be therapeutic against obesity, targeting lipase with an IC_{50} value lower than that of vibralactone (18).

While known biosynthetic pathways that utilize a β-lactone intermediate are few, there are many documented uses for β-lactones in chemical synthesis. The high ring strain and two sites for attack at the C2 and C4 position allow for multiple different transformations. These include electrophilic addition using an enolate intermediate, decarboxylation of the ring to an alkene, Lewis acid catalyzed rearrangements, polymerization, nucleophilic attack, and conversion to β-lactams, a popular class of antibiotics, or γ-lactams, which have the ability to undergo many more transformations (1, 3, 4, 11). Further research into novel β-lactones not only increases the types of β-lactones that can be created, but also increases the possibility of discovering other chemically important derivatives.

Current Understanding of β-Lactone Biosynthesis

A review by Robinson *et al.* details five main classes of β-lactones based upon the substrates from which they are derived (4). While there are some that do not fit within these categories, the majority do. These β-lactones are thought to arise from either 1) amino acids, 2) fatty acids, 3) polyketides, 4) hybrid polyketide synthetase/nonribosomal peptide synthetase (PKS/NRPS) pathways, or 5) terpenes (3, 4). In addition, there are multiple methods in nature to synthesize β-

lactones, with perhaps more yet to be discovered. Currently, there are four mechanisms for β-lactone synthesis that have been published (4, 6, 19–21). Two have been shown through *in vitro* studies, and two others have been proposed based on biochemical evidence. The first mechanism, discovered in January of 2017, relies on a β-hydroxy acid intermediate activated by AMP (Figure 1-2A) (6). The second also relies on a β-hydroxy acid intermediate, though this is activated through the thioesterase (TE) domain of a non-ribosomal peptide synthetase (NRPS). This mechanism was discovered in the NRPS ObiF, originating in the pathway that produces obafluorin (Figure 1-2B) (21). The biosynthesis of salinosporamide A has been proposed to begin with a β-hydroxy acid intermediate attached to a peptidyl carrier protein (PCP) of SalB, another NRPS protein (Figure 1-2C) (20). Finally, vibralactone has been thought to arise from a 2-(3*H*)-oxepinone intermediate but the enzyme and the actual mechanism that produces the β-lactone is unknown (Figure 1-2D) (19).

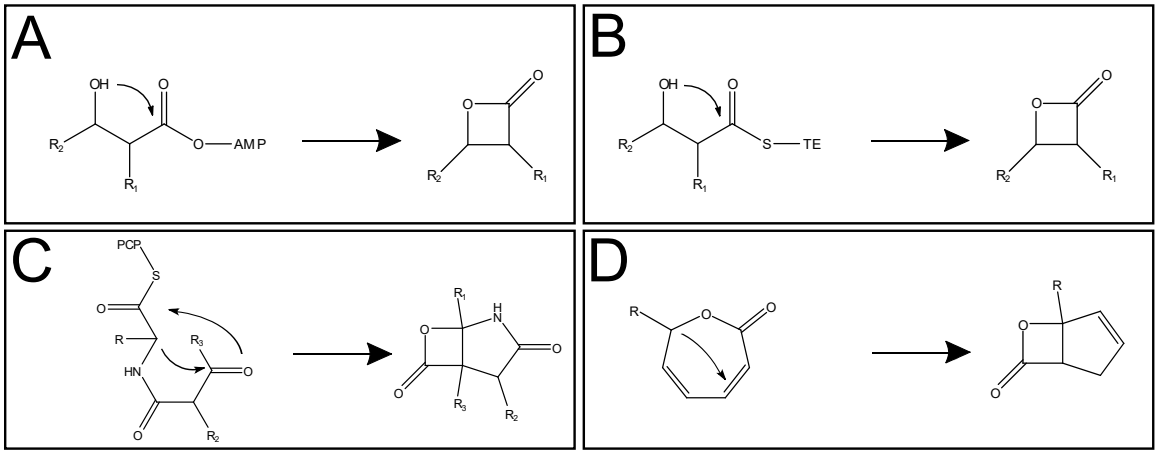


Figure 1-2. Mechanism of β-lactone biosynthesis. A.) AMP activation of a β-hydroxy acid intermediate. B.) Activation of a β-hydroxy acid by the TE domain of an NRPS. C.) Activation of a β-hydroxy acid by PCP. D.) Proposal for lactonization of the 2-(3*H*)-oxepinone intermediate.

One issue with discovery and isolation of β-lactone natural products, and more generally other potential antibiotics, is the existence of “cryptic” pathways

for antibiotic biosynthesis that may only be activated under certain conditions (22). If an antibiotic biosynthesis pathway is discovered through genome mining or other techniques but is considered “cryptic”, meaning the biosynthesis pathway is not expressed, there are a variety of ways to try to stimulate production of the antibiotic. Beyond the more traditional methods of stress and changing culture conditions, engineering can be a way to either induce production in the natural host or reconstruct the pathway in an organism more favorable for fermentation (22). Additionally, there are limitations on what structures can be found in nature. While there are likely more β-lactone biosynthesis pathways to be discovered than the ones discussed in this section, there may be structures with higher therapeutic potential that do not exist as natural products.

Chemical Synthesis of β-Lactones

While the high ring strain of the β-lactone make it a desirable synthon for other downstream products, chemical synthesis of the β-lactone can be difficult as it is highly unstable. Still, there are numerous methods for the synthesis of this class of compounds (1, 8, 9, 11, 23, 24). These methods can be laborious with low yield. For example, a recently published total synthesis of the natural product caulolactone A required 10 steps and had a 26% overall yield (11). Low yields and costly chemical syntheses create a barrier for the development of novel β-lactones that cannot be found in nature. This can be an issue when developing a promising synthesized β-lactone product for use as a therapeutic. One example of this challenge is found in a study by Richardson *et al.* that synthesized nineteen β-lactones that were structurally similar to Orlistat (lipstatin), yet had increased activity against fatty acid synthase, responsible for tumor cell survival in some cancers (8). These β-lactones had an increase in tumor cell death compared to Orlistat, yet synthesis was completed in six steps with a 20 percent yield from the starting *n*-octanal (8). A biosynthetic approach may be able to improve yield, especially for β-lactone structures that are already similar to known biosynthetic natural products. In another example, β-lactone structures

were developed that had a dual action against both the fatty acid synthase and proteasome, both of which are targets for anticancer drugs (9). Using knowledge from both belactosin C, a proteasome inhibitor, and lipstatin, a known fatty acid synthase inhibitor, the authors of the article were able to develop novel structures that could work on both (9). The synthetic β-lactones developed in the study have a chance to avoid resistance due to the dual mode of action. Therefore, research into novel β-lactones could not only improve yield, but also lead to more effective therapies with a lower risk of resistance.

Potential for Biosynthesis of Novel β-Lactones

One way to overcome the limitations of both natural isolation and chemical synthesis is through engineered biosynthesis of novel β-lactone. Enzymatic biosynthesis pathways in cultured organisms generally result in higher yields of product with little environmental waste compared to chemical synthesis. Additionally, engineering of biosynthetic pathways and improved fermentation methods can lead to novel biosynthetic methods and an overall improved yield when compared to natural biosynthesis of the desired product. Strides in improved fermentation of β-lactones – namely with tetrahydrolipstatin, the precursor to the antiobesity drug lipstatin (5) – have been made, and studies have demonstrated the potential for engineering novel biosynthesis pathways in the creation of valuable compounds (25, 26).

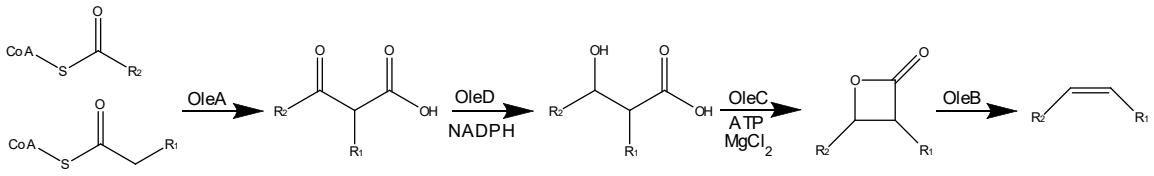
Olefin Biosynthesis Pathway

The β-lactone pathway, the focus of this thesis, is derived from the Olefin biosynthesis pathway, OleABCD. This pathway has been studied for the creation of olefinic hydrocarbons that make up portions of microbial membranes. However, it has been only recently that the first β-lactone synthetase enzyme was characterized (6). Owing to the instability of the β-lactone ring, it had been believed that OleC formed the final Olefin structure. However, further investigation revealed that OleC was in fact forming a β-lactone, while OleB collapsed the ring into the olefin(6, 27). This discovery was foundational for the

work presented in this thesis. Because of the extensive characterization of the Ole pathway in *Xanthomonas campestris,* as well as its homology to other known β-lactone pathways, this pathway was used as the model for the work presented in this thesis.

The pathway consists of four proteins and uses fatty acyl-CoA substrates to create olefin (Figure 1-3a). OleA is the first enzyme in the pathway. It is a non-decarboxylative thiolase that catalyzes the Claisen condensation between two fatty acyl-CoAs to form a β-keto acid (29). OleD is a 2-alkyl-3-ketoalkonic acid reductase that uses NADPH as a cofactor to reduce the β-keto acid to a β-hydroxy acid (30). OleC, the β-lactone synthetase is the next enzyme in the pathway and forms the β-lactone, using ATP and $MgCl_2$ as cofactors (6). Finally, OleB is a β-lactone decarboxylase that collapses the ring structure into the resulting olefin (27).

A



B

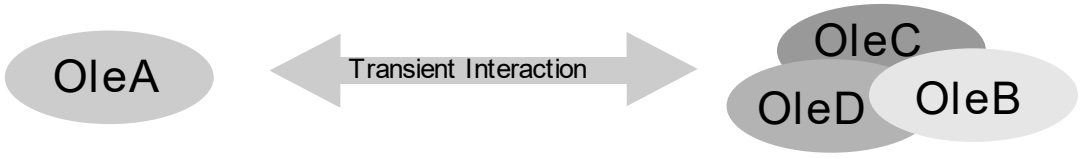


Figure 1-3. The OleABCD pathway. A.) A schematic of OleABCD reactions. B.) A diagram of the OleBCD complex with OleA only transiently interacting.

The enzymes are thought to interact with the membrane, with OleBCD existing as a complex, and OleA only transiently interacting with the OleBCD complex (Figure 1-3b) (28). It was unknown why OleA was not an integral part of

the complex. Preliminary data I gathered during the duration of this thesis indicated OleA was inhibited by the β-lactones, presumably through an attack of the ring by the active site cysteine (Supplemental Figure 1-S1a). Further experients showed that when OleA, OleB and OleC were incubated in solution with myristoyl-CoA, the resulting β-lactone produced after a four-hour incubation was able to inhibit OleA (Supplemental Figure 1-S1b). Additionally, an experiment was performed in collaboration with Dr. Robinson that included incubation of OleA with β-lactone and analyzed by MALDI-TOF (Supplemental Figure 1-S2). This showed a shift in the mass of the enzyme corresponding to the β-lactone, which indicates that the inhibition of OleA may be covalent. Together, the data from these experiments suggest that OleA may transiently interact with the OleBCD complex to avoid inhibition.

A Look at the OleA Mechanism

OleA is a thiolase that carries out a non-decarboxylative Claisen condensation, combining two acyl-CoA substrates and forming a β-keto acid (15, 29). OleA was found to proceed in three distinct steps through a ping-pong mechanism (Figure 1-4) (29).

The first step involves transesterification of the acyl-chain from the first acyl-CoA substrate to the active site cysteine (Cys143). Crystal structures of OleA with substrate as well as inhibitors such as cerulenin show that the acyl chain lies within one of two channels, defined as channel “A”(31, 32). The correct positioning of the chain is thought to be determined by the presence of one of two oxyanion holes created by His285 and the side chain of Asn315 (33). The CoA portion of the acyl-CoA has been observed in the crystal structures to lie outside the OleA protein, with the pantetheine portion of the CoA residing in a channel known as the “P” or “pantetheine” channel. As the cysteine attacks the carbonyl of the acyl-CoA thioester, the CoA is released, allowing the now formed enzyme substrate complex to accept a second acyl-CoA substrate.

The alkyl chain of the second acyl-CoA substrate is thought to reside in the second channel or channel “B”. It is at this point the second step in the OleA

mechanism – the Claisen condensation – can begin (32). Prior to this thesis work, the directionality of the Claisen condensation was unknown, but Chapter 4 delineates how Glu117 acts as a general base to deprotonate the α-carbon of the first substrate alkyl chain(34, 35). This carbocation is then primed for attack of the carbonyl of the second alkyl chain, completing the Claisen condensation and forming the β-keto acid. Finally, the third step of the reaction involves release of the β-keto acid from the active site, freeing the enzyme for another round of catalysis.

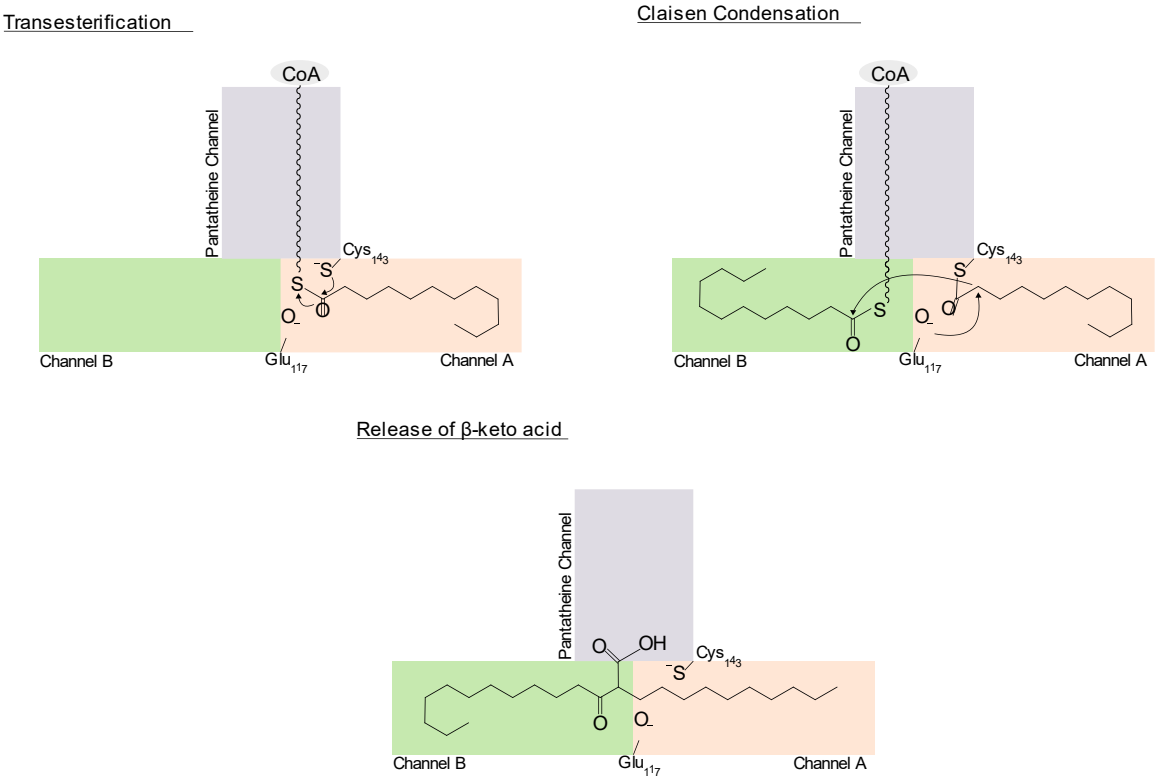


Figure 1-4. Schematic of the OleA mechanism.

Potential for "Plug and Play" Biosynthesis of novel β-lactones

It was hypothesized that within the model pathway, OleABCD from *X. campestris,* it is possible to use a "plug and play" approach to biosynthesis of novel β-lactone natural products. It has been shown that homologous enzymes in the olefin pathway can be substituted without detriment to the microorganism to

create novel products (36, 37). Our work has shown that homologs of this pathway, as well as OleA homologs have been found in a wide number of organisms (3, 4, 36, 38, 39). One example of how this approach would work is through the use of enzymes from the pathway for the biosynthesis of nocardiolactone, Nlt. Contributions I made to the reconstruction of the Nlt pathway were published in the Journal of Biological Chemistry by Robinson et al (39). The NltABCD enzymes are homologous to those in the the model *X. campestris* Ole pathway (Figure 1-5A), except that it lacks an OleB β-lactone decarboxylase homolog that collapses the β-lactone ring into the final olefin product. Instead the NltAB together make up a heterodimer that is homologous to the OleA homodimer, which is similar to the LstAB heterodimer used in lipstatin bioysnthesis (39, 40). We substituted the functionally equivalent OleA for the less stable NltAB heterodimer in order to generate more product.

In a parallel unpublished experiment, I showed that a substitution of OleD for NltD in the Nlt pathway resulted in a switch from the major product of *trans* β-lactone for *c* β-lactone, seen as the more stable olefin product on GC-MS (Figure 1-5C). OleD has been shown to have a specificity for the *syn* β-keto acid, specifically the (2R,3S) isomer (30). While the isomeric substrate specificity of NltD has not yet been published, the nocardiolactone structure was isolated as a *trans β*-lactone, so it stands to reason that of the four enantiomers of β-keto acid, the NltD would accept one of the two *anti iso*mers (41, 42). Both OleC and NltC β-lactone synthetase homologs did not distinguish between diastereomers of the intermediate. Therefore, we were able to demonstrate that by switching one enzyme in the pathway we could control the stereochemistry of the resulting β-lactone. This approach of substituting homologs with desired specificities could potentially be used to introduce larger changes to the overall chemical structure of the β-lactone.

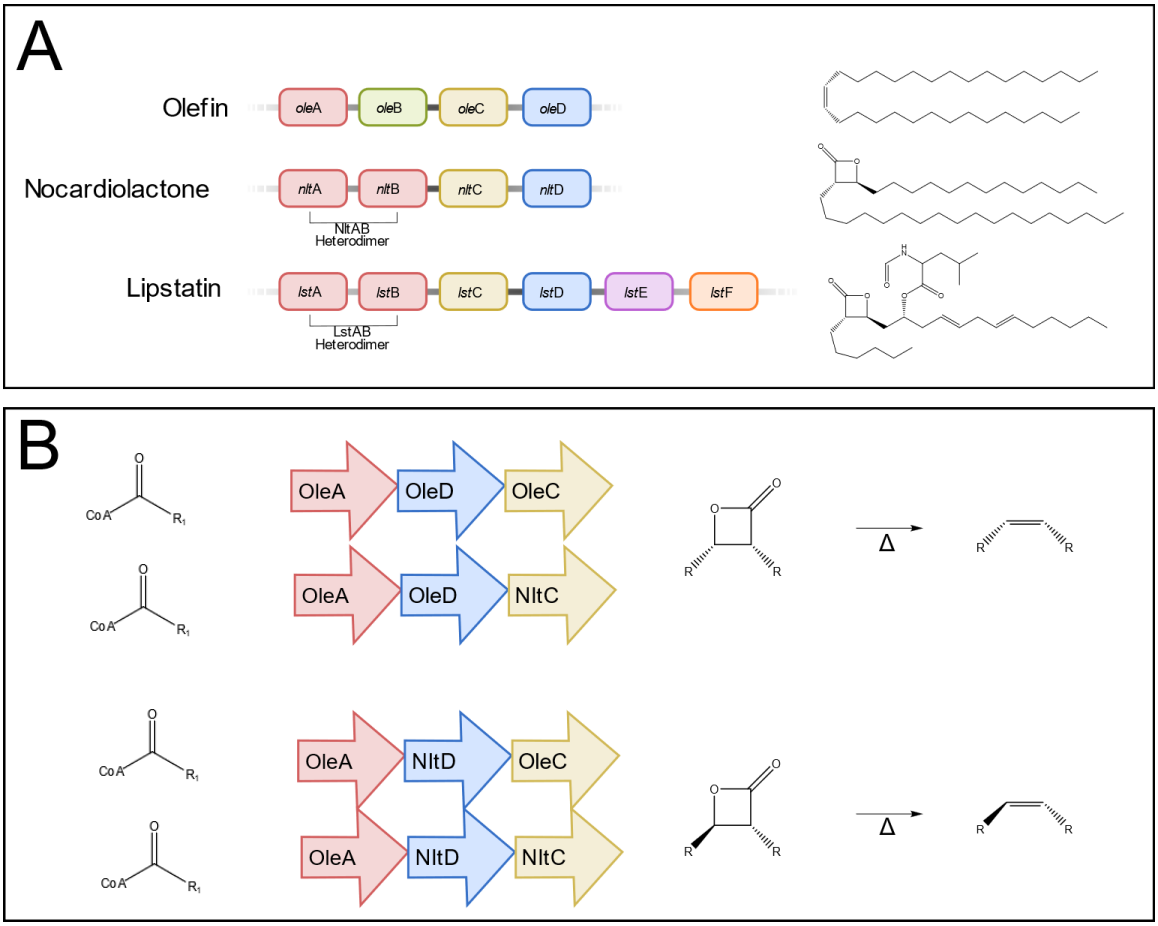


Figure 1-5. The homologous Ole, Nlt, and Lst pathways. A.) Comparison of *ole, nlt,* and *lst* gene clusters and their homology. B.) Outcomes of four *in vitro* pathways tested. Fatty acyl-CoA was incubated with either OleA, OleD or NltD, and OleC or NltC to produce β-lactone, which was decarboxylated in the heat of the GC to an olefin. The majority product for reaction mixtures containing OleD was *syn*, while the majority product for mixtures containing NltD was *anti*. Swapping OleC for NltC had no effect on the structure of the product. For more information see Supplementary figure 1-S3.

Some β-lactone pathways contain additional enzymes that modify the β-lactone structure. The lipstatin (Lst) pathway is a well characterized example of β-lactone natural product pathway homologous to the Nlt and Ole pathways (12), that includes LstE and LstF. These enzymes are thought to attach an N-

formylated leucine residue to the β-lactone structure (Figure 1-5A) (12), which could allow for the production of many variations on the β-lactone structure.

OleA – A Major Focus of this Thesis

Swapping Ole enzymes has led to novel products. OleA, the first enzyme in the pathway acts as the gate keeper, determining the overall structure. Goblirsh *et al* demonstrated how substitution of an OleA homolog from *Stenotrophomonas maltophilia* into *Shewanella oneidensis* MR-1 could change the profile of the resulting olefins (36). Because of this, OleA became a major focus of this thesis research as it could have the most dramatic effect on the structure of the β-lactone products. In addition, a recent grant by the Joint Genome Institute provided the ability to probe the natural diversity of potential OleA homologs (discussed in Ch 2 and 3), by providing us with around 5000 kb worth of DNA, which we used to produce distinct OleA homologs and *X. campestris* OleA mutants.

Prior to finding that OleA catalyzed the hydrolysis of *p*-nitrophenyl esters, OleA activity was followed with lengthy stop time assays. While this was sufficient for prior mechanistic work, it would be very difficult to use these techniques for the large number of putative OleA homologs we received from the JGI institute for this research. Thus, it was necessary to develop a continuous and rapid assay to test multiple OleA enzymes at once. My chance discovery of the hydrolysis of *p*-nitrophenyl esters by OleA provided the opportunity to work with these constructs within the time frame of a doctoral thesis through the development of a rapid kinetic assay detailed in Chapter 2 of this thesis (43).

Thesis Prelude and Author Contributions

This work aims to be a step towards the biosynthesis of novel β-lactone therapies through the research of a model β-lactone biosynthetic pathway, specifically focusing on the enzyme that determines the overall structure.

In chapter 2, I develop a novel assay using a colorimetric analog, the *p*-nitrophenyl ester, and show how it can be used to probe diversity of OleA

homologs. I am responsible for the discovery, design, application, and verification of the assay, along with aid from my undergraduate mentee, Mandkhai Molomjamts. Dr. Serina Robinson performed all bioinformatic analysis. Dr. Robinson and I collaborated to choose the 72 putative OleA homologs, with Dr. Robinson performing the analysis to choose the 27 OleAs, and I choosing the 43 manually. Dr. Lawrence Wackett oversaw all research and writing and mentored me through design of the experiment. All authors edited and contributed to the published manuscript.

In collaboration with Dr. Robinson, an experienced bioinformaticist, we show how this method can be further used to probe substrate specificity of these homologs and develop a model that can be used in future engineering research in chapter 3. Dr. Jack Richman synthesized many of the *p*-nitrophenyl esters used in this study. I oversaw the application of the whole-cell *p*-nitrophenyl ester colorimetric assay for this work and worked alongside Dr. Robinson and Dr. Kelly Aukema to collect all the data used in this study. Dr. Robinson was responsible for much of the data analysis, machine-learning model development, as well as development of the web application. The manuscript was primarily written by Dr. Robinson. Dr. Wackett oversaw all research, writing, and provided guidance at all stages of research and journal article preparation. All authors edited and contributed to the published manuscript.

I then use the second half of this thesis to investigate the biotechnological aspect of *p*-nitrophenyl esters and use them to both clarify the mechanism of OleA and to develop novel β-keto acid in chapter 4. Here I am responsible for the design of the experiments, and perform all *in vitro* studies, aided by my undergraduate mentee, Troy Biernath. Dr. Aukema assisted in gathering data for multiple experiments, as well as assisted in mentoring Troy Biernath for a portion of the time dedicated to this research. Lambross Tassoulas provided expertise in protein modeling and analyzed published OleA crystal structures. Dr. Richman is responsible for synthesis and instruction for handling of all *p*-nitrophenyl esters not commercially available that were used in this assay. Dr. Wackett oversaw

research, and aided in writing the manuscript, as well as mentorship. All authors edited and contributed to the manuscript.

Finally, in chapter 5, I demonstrate how thiolysis of *p*-nitrophenyl esters by CoA can be exploited in order to generate novel β-keto acid. I also demonstrate how cetyltrimethylammonium bromide (CTAB) can be used to improve the yield of OleA and provide some evidence into the mechanism of this improvement. I am responsible for design of the experiments, as well as all studies, with the aid of my undergraduate mentee, Troy Biernath. I am also responsible for the preparation of this chapter. Dr. Richman provided synthetic *p*-nitrophenyl esters that are not commercially available. Dr. Wackett oversaw the experiments and provided guidance for developing the project. He has also provided some mentorship in the preparation of this chapter.

The research contained in this thesis represents a contribution to the study of biosynthetic pathways in microorganisms. This work can be expanded in future studies for the development of non-natural β-lactone, and techniques used in this work may have applications in the study of other industrially relevant compounds.

CHAPTER 2 – *In Vivo* Assay Reveals Microbial OleA Thiolases Initiating Hydrocarbon and β-Lactone Biosynthesis

---

"Sometimes the most scenic roads in life are the detours you didn't mean to take." – Angela N. Blount (Once Upon Ever After)

This chapter is reprinted with permission from American Society for Microbiology *mBio.*

mBio, 2020, 11(2), e00111-20
DOI: 10.1128/mBio.00111-20

In Vivo Assay Reveals Microbial OleA Thiolases Initiating Hydrocarbon and β-Lactone Biosynthesis

Megan D. Smith, Serina L. Robinson, Mandkhai Molomjamts, and Lawrence P. Wackett

Summary

OleA, a member of the thiolase superfamily, is known to catalyze the Claisen condensation of long-chain acyl coenzyme A (acyl-CoA) substrates, initiating metabolic pathways in bacteria for the production of membrane lipids and β-lactone natural products. OleA homologs are found in diverse bacterial phyla, but to date, only one homodimeric OleA has been successfully purified to homogeneity and characterized in vitro. A major impediment for the identification of new OleA enzymes has been protein instability and time-consuming in vitro assays. Here, we developed a bioinformatic pipeline to identify OleA homologs and a new rapid assay to screen OleA enzyme activity in vivo and map their

taxonomic diversity. The screen is based on the discovery that OleA displayed surprisingly high rates of p-nitrophenyl ester hydrolysis, an activity not shared by other thiolases, including FabH. The high rates allowed activity to be determined *in vitro* and with heterologously expressed OleA *in vivo* via the release of the yellow *p*-nitrophenol product. Seventy-four putative *oleA* genes identified in the genomes of diverse bacteria were heterologously expressed in *Escherichia coli*, and 25 showed activity with *p*-nitrophenyl esters. The OleA proteins tested were encoded in variable genomic contexts from seven different phyla and are predicted to function in distinct membrane lipid and β-lactone natural product metabolic pathways. This study highlights the diversity of unstudied OleA proteins and presents a rapid method for their identification and characterization.

Importance

Microbially produced β-lactones are found in antibiotic, antitumor, and antiobesity drugs. Long-chain olefinic membrane hydrocarbons have potential utility as fuels and specialty chemicals. The metabolic pathway to both end products share bacterial enzymes denoted as OleA, OleC, and OleD that transform acyl-CoA cellular intermediates into β-lactones. Bacteria producing membrane hydrocarbons via the Ole pathway additionally express a β-lactone decarboxylase, OleB. Both β-lactone and olefin biosynthesis pathways are initiated by OleA enzymes that define the overall structure of the final product. There is currently very limited information on OleA enzymes apart from the single representative from *X. campestris*. In this study, bioinformatic analysis identified hundreds of new, putative OleA proteins, 74 proteins were screened via a rapid whole-cell method, leading to the identification of 25 stably expressed OleA proteins representing seven bacteria phyla.

Introduction

Bacteria produce different fatty acid derivatives that serve a structural role in cell membranes or that are secreted as natural products for intracellular signaling and competition (44–46). In most cases, those biosynthetic pathways

are independent. Recently, a set of proteins denoted as OleA, OleC, and OleD showed overlapping function, being important in the microbial production of membrane hydrocarbons and β-lactone natural products (Figure 2-1) (6).

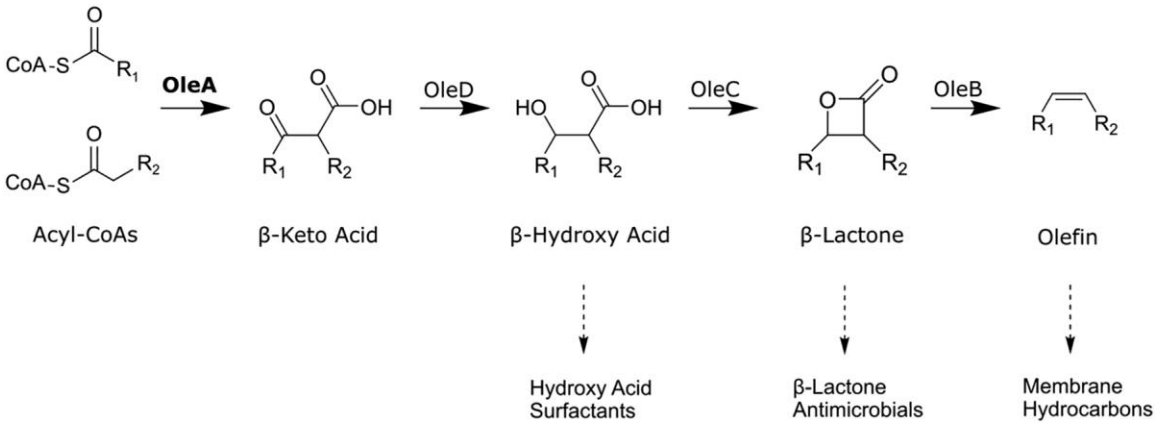


Figure 2-1. Bacterial metabolism using OleA proteins to initiate the pathways. Final products are from right to left, olefinic hydrocarbons that are components of membranes, β-lactones that serve as natural product enzyme inhibitors, and functionalized hydroxy acids that are produced by some bacteria as surfactants that act to solubilize hydrophobic substrates.

Previously, we studied Ole enzymes in the model pathway found in the plant-pathogenic bacterium *X. campestris*. In the *X. campestris* pathway, the OleA, OleC, and OleD enzymes are coexpressed with OleB, which catalyzes an unprecedented enzymatic decarboxylation of β-lactones (27) to generate hydrophobic membrane olefins (Figure 2-1). However, other bacteria harbor gene clusters encoding homologous Ole proteins but lack an *oleB* gene. Bacteria with this *oleADC* gene cluster create secreted β-lactone natural products, some of which are found to have antibiotic, anticancer, or antiobesity properties in medical testing (4). For example, salinosporamide A, a β-lactone natural product, is currently in phase three clinical trials for the treatment of glioblastoma (7). Lipstatin, produced by various *Streptomyces* spp., is hydrogenated industrially to make the FDA-approved antiobesity drug tetrahydrolipstatin (5, 12).

Both β-lactone and olefin pathways start with a nondecarboxylative Claisen condensation of acyl coenzyme A (acyl-CoA) precursors by OleA (Figure 2-1), a member of the thiolase superfamily (29, 36, 40). Many thiolase enzymes catalyze carbon-carbon bond formation (47). They function in fatty acid, hydrocarbon, and natural product biosynthesis. As such, there is significant interest in identifying and reengineering these enzymes for biotechnological purposes (48). Our understanding thus far is derived almost exclusively from mechanistic and X-ray crystallographic studies with OleA from *X. campestris* (31–34) and a study on hydrocarbon biosynthesis in *Micrococcus luteus* (16). *X. campestris* OleA catalyzes the condensation of acyl-CoA substrates with C_{10}-C_{16} acyl chains and produces long-chain hydrocarbons via deoxygenation reactions catalyzed by OleC and OleB proteins. Two distant homologs of OleA are LstA and LstB that together form a heterodimer in solution and catalyze the condensation of (3S,5Z,8Z)-3-hydroxytetradeca-5,8-dienoyl-CoA and octyl-CoA to produce the backbone of the β-lactone natural product, lipstatin (40).

The thiolase superfamily largely contains enzymes that join a short carbon chain to a growing chain, for making fatty acids or polyketide natural products (47). The singular characterized OleA differs from most thiolases in condensing two acyl chains ranging from C_8 to C_{16} (29). However, the divergence in OleA and other thiolase sequences, as low as ~17 to 30% amino acid identity, makes it currently impractical to predict function from sequence alone. To date, demonstrating a thiolase protein to be an OleA enzyme has required purifying the protein and carrying out a time-consuming assay. The reported assay for OleA activity is discontinuous and requires solvent extraction, gas chromatography, and calibration with authentic standard compounds that are not commercially available and require multistep syntheses (29, 32, 47). During the assay, the physiological β-keto acid product undergoes spontaneous decarboxylation, and the resultant ketone is quantified as a surrogate for the keto acid. This assay, as well as the poor solubility of these proteins, has precluded purification and characterization of OleA. Consequently, our current understanding of the biology and chemistry of OleA has largely been confined to the protein from *X.*

campestris, although *in vivo* and bioinformatic analyses suggest that diverse bacterial strains produce OleA enzymes to make a wide range of different products (4, 36). The substrate specificity of OleA thus demonstrates the structures of the downstream products (Figure 2-1). In this context, the identification and characterization of additional OleA enzymes provide the key to diverse natural products and membrane components generated by Ole proteins.

In the current work, we observed that the *X. campestris* and other OleA proteins will accept *p*-nitrophenyl alkanoates and catalyze a hydrolysis reaction to release the yellow product *p*-nitrophenol (*p*-NP). The tested thiolases FabH and Pks13 did not react, and the OleA reaction rates were surprisingly comparable to rates observed with lipases and other hydrolytic enzymes assayed with *p*-NP esters (49–52). This allowed the utility of *p*-NP ester reactivity to quantify OleA activity both *in vitro* and *in vivo*. The latter method, described here, served to rapidly screen 74 OleA homologs and identify and characterize new OleA proteins.

Results

OleA enzymes are found in taxonomically diverse bacteria. Despite the diversity of products made by Ole proteins (4, 36), previous efforts to purify five other OleA proteins were unsuccessful (29). This suggested that many OleA proteins are not amenable to purification and expression in heterologous hosts. In this context, genome sequences were analyzed here to identify divergent *oleA* genes that might produce stable and active OleA proteins when expressed heterologously in *Escherichia coli*.

In a broad screening effort, divergent OleA protein homologs were identified in 17 different taxonomic classes of bacteria, including *Chlamydiae*, *Clostridia*, *Opitutae*, and *Oligoflexia* (Figure 2-2). Sequence cluster representatives belonging to 74 different organisms were selected as described in Materials and Methods to do the following: (i) define the sequence signature of true OleA proteins within the thiolase superfamily, (ii) screen for OleA proteins likely to express in active form in *E. coli*, and (iii) identify new OleA proteins that

aid in determining the structural diversity of products made by OleA-initiated metabolic pathways. As is shown in Figure 2-2, there are a limited number of characterized natural products that are associated with OleA proteins. Most of the proteins were annotated in GenBank as 3-oxo-acyl-acyl carrier protein (ACP) synthase III proteins. The amino acid sequence identities of the proteins screened here ranged from 24.3 to 87.9% to the *X. campestris* OleA amino acid sequence (see Table 2-S1 in the supplemental material).

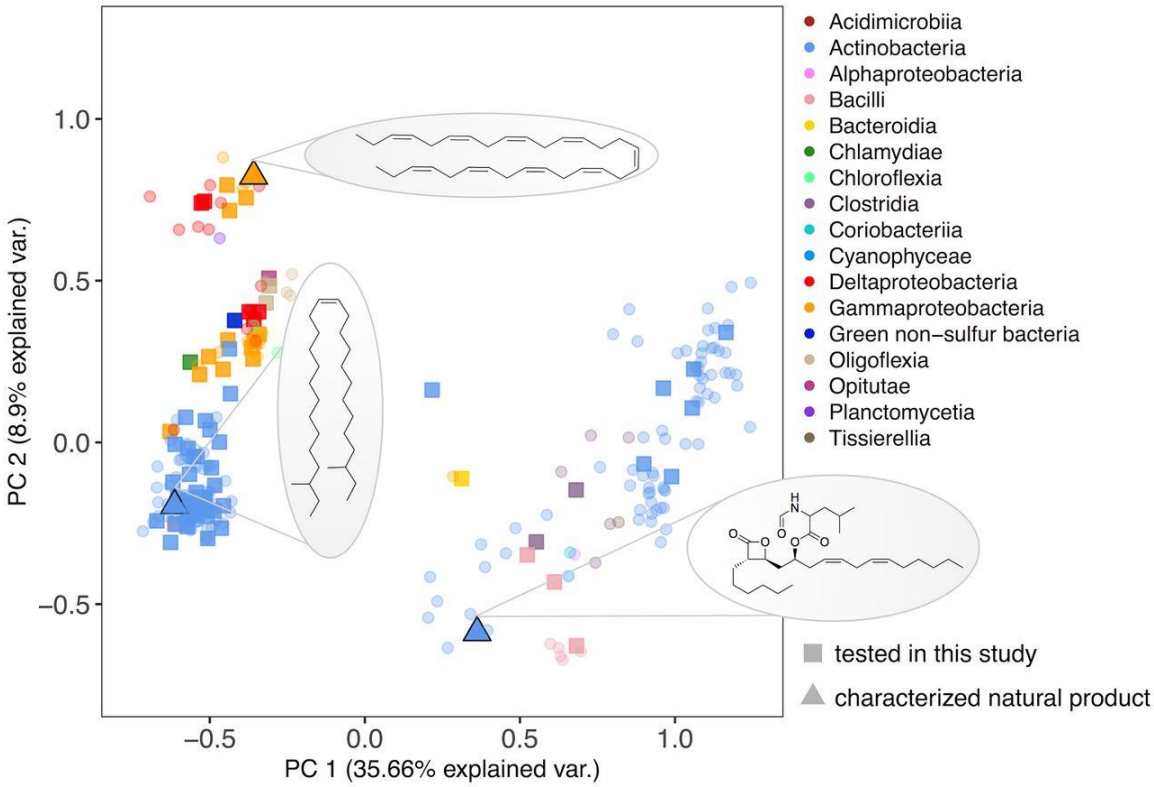


Figure 2-2. Sequence space of OleA proteins visualized using principal-coordinate analysis. The Whelan and Goldman distance matrix was used to calculate dissimilarity between full-length OleA amino acid sequences. Sequences are colored by taxonomic class. Circles represent 235 *oleA* homologs with flanking *oleCD* genes selected as cluster representatives from the thiolase superfamily (PF08545). Squares correspond to the 74 *oleA* genes that were synthesized and tested in this study. Triangles indicate OleA sequences for

organisms with a structurally characterized Ole pathway product, a β-lactone or olefinic hydrocarbons, as shown in the ovals. var., variation.

Development of a colorimetric assay for OleA *in vitro*. OleA proteins cannot be assayed in crude extracts with acyl-CoA substrates effectively due to interfering thioesterase activity, so studies thus far have been confined to *in vitro* experiments with purified enzymes (Figure 2-3A). Over the course of *in vitro* coincubations with *X. campestris* OleA and porcine lipase, we observed unexpectedly high rates of hydrolysis of the lipase *p*-nitrophenyl ester substrate. Subsequently, it was determined that OleA alone catalyzed rapid color formation with *p*-nitrophenyl laurate. Indeed, when we compared rates of *p*-nitrophenyl ester hydrolysis of OleA compared to known hydrolases, OleA showed equivalent or greater rates (Supplemental Table 2-S2). OleA was shown to produce *p*-nitrophenol and a fatty acid via UV-visible (UV-Vis) spectroscopy and gas chromatography (see Supplemental Figure 2-S1 in the supplemental material).

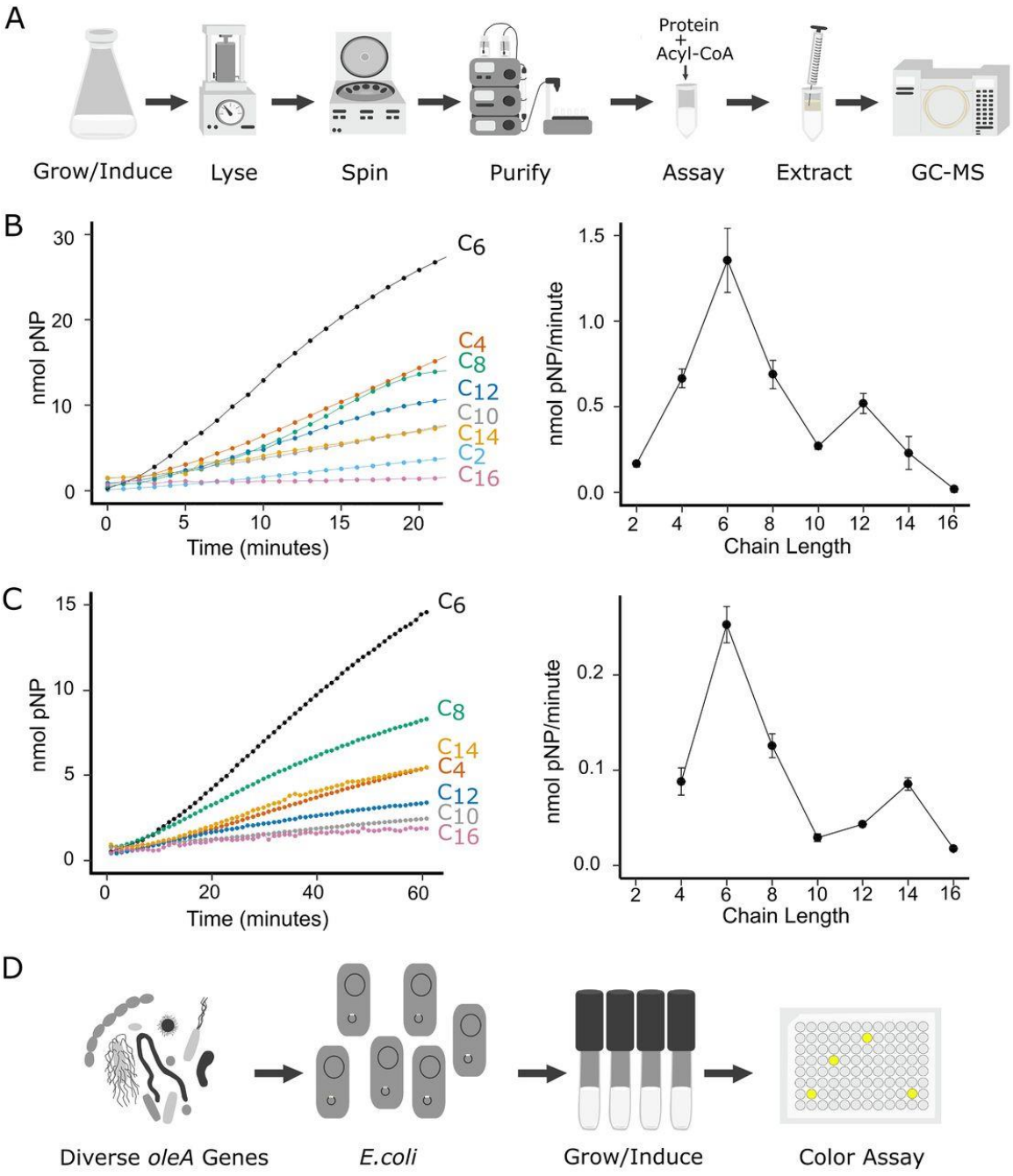


Figure 2-3. Assay schematic and data illustrating rapid assay compared to established methods to assay OleA activity. (A) The established OleA assay required purified proteins and expensive acyl-CoA substrates, reaction extractions, and gas chromatography-mass spectrometry (GC-MS) analysis of reaction products. (B) Assays conducted with purified OleA from *X. campestris* using *p*-nitrophenyl acyl substrates of different chain lengths. (C) Whole-cell

assays conducted with E. coli cells expressing *X. campestris* OleA and using *p*-nitrophenyl acyl substrates of different chain lengths. In panels B and C, rate curves are shown on the left and activity as a function of acyl chain length is shown on the right. (D) The rapid assay described in this paper screens diverse *oleA* genes, expressed recombinantly in *E. coli* cells using a colorimetric assay in microtiter well plates.

Since *p*-nitrophenyl laurate has low water solubility and *X. campestris* OleA is promiscuous with respect to acyl-CoA chain length (C_{10} to C_{16}), other *p*-nitrophenyl alkyl ester chain lengths were investigated. The reaction was observed to proceed similarly in microtiter wells or individual cuvettes, so microtiter plates were used in all subsequent experiments. Purified OleA was incubated individually with *p*-nitrophenyl esters containing all even-number carbon chain lengths ranging from C_2 to C_{16}, as shown in Figure 2-3B. The reaction time course with C_2 and C_4 acyl chain lengths was linear from the first time points taken at less than 1 min. Longer chain length esters showed an initial lag phase before displaying a linear increase for 10 min or more. This has been observed previously in lipase assays and attributed to poor water solubility and longer dispersal times for the longer chain length esters (53). Maximum activity was observed with C_6, followed by a linear decrease in activity with increasing chain length with the exception of a second smaller spike in activity at C_{12} (Figure 2-3B). Given the poor solubility of the substrates, steady-state kinetic parameters could not be reliably measured, as previously discussed by Guthrie for *p*-nitrophenyl ester reactions with lipases (53). However, the observed rate with *p*-nitrophenyl hexanoate of 350 nmol per min per mg of protein was substantial, suggesting a potential for whole-cell screening of E. coli cells heterologously expressing different OleA proteins.

Rapid *in vivo* screen for OleA. On the basis of published reports (53–55), we expected that the long-chain *p*-nitrophenyl acyl esters would generally not enter E. coli cells without a permeabilizing agent. Following the general protocol for a

high-throughput assay for oxygenases that used *p*-nitrophenyl ethers, polymyxin B sulfate was used as an E. coli cell permeabilizer (53, 54, 56). With this additive, rates for hydrolysis of C_4 to C_{16} chain length *p*-nitrophenyl esters were significantly higher than the background, allowing a determination of OleA activity (Figure 2-3C). Reaction rates with permeabilized cells per unit OleA protein estimated for the recombinant cells closely resembled the rates observed previously with purified enzyme. Moreover, the relative rates of different chain lengths followed a nearly identical pattern for purified enzyme (Figure 2-3B) and permeabilized E. coli cells expressing OleA (Figure 2-3C).

While the assay above is amenable to rapid screening, albeit with a 2-h preincubation to permeabilize cells, it would be faster and much more convenient if substrate could be added directly to cell suspensions without pretreatment. To test this, we used the highest-turnover substrate with moderate water solubility, *p*-nitrophenyl hexanoate. By direct addition of *p*-nitrophenyl hexanoate to cell suspensions in microtiter wells, yellow product formation over time could be monitored readily and at approximately 67% of the rate determined with permeabilized cells. In light of these results, the extensive purification and assay procedure (Figure 2-3A) was replaced with a direct substrate drop-in assay with 96 reactions conducted simultaneously in microtiter plates (Figure 2-3D). This assay was instrumental in identifying new, expressible, and active OleA enzymes.

Screening of diverse bacterial thiolase proteins for identifying and obtaining new OleA proteins. Seventy-four recombinant E. coli strains containing genes for putative OleA proteins were screened *in vivo* using direct addition of *p*-nitrophenyl hexanoate (Figure 2-4). Hydrolysis rates for recombinant cells were normalized against E. coli containing an empty vector, and rates for positively increasing slopes were recorded (Supplemental Figure 2-S2). In order to remove potential false-positive results, only those with a rate above the mean activity were recorded as active (Supplemental Figure 2-S3). The final compiled rates for all proteins screened across three biological

replicates (Supplemental Table 2-S3) were mapped onto a phylogenetic tree of OleA sequences (Figure 2-4). Phylogenetic analysis revealed two distinct clades of actinobacterial OleA sequences with one clade displaying significant activity and the other displaying only weak activity in one homolog (Figure 2-4). In total, excluding our positive-control OleA from *X. campestris,* 25 OleA proteins were found to be active, representing ~35% of the proteins screened for activity. This proportion is an increase to our previous experience with OleA proteins in which one in five could be produced as active enzymes in *E coli* (29).

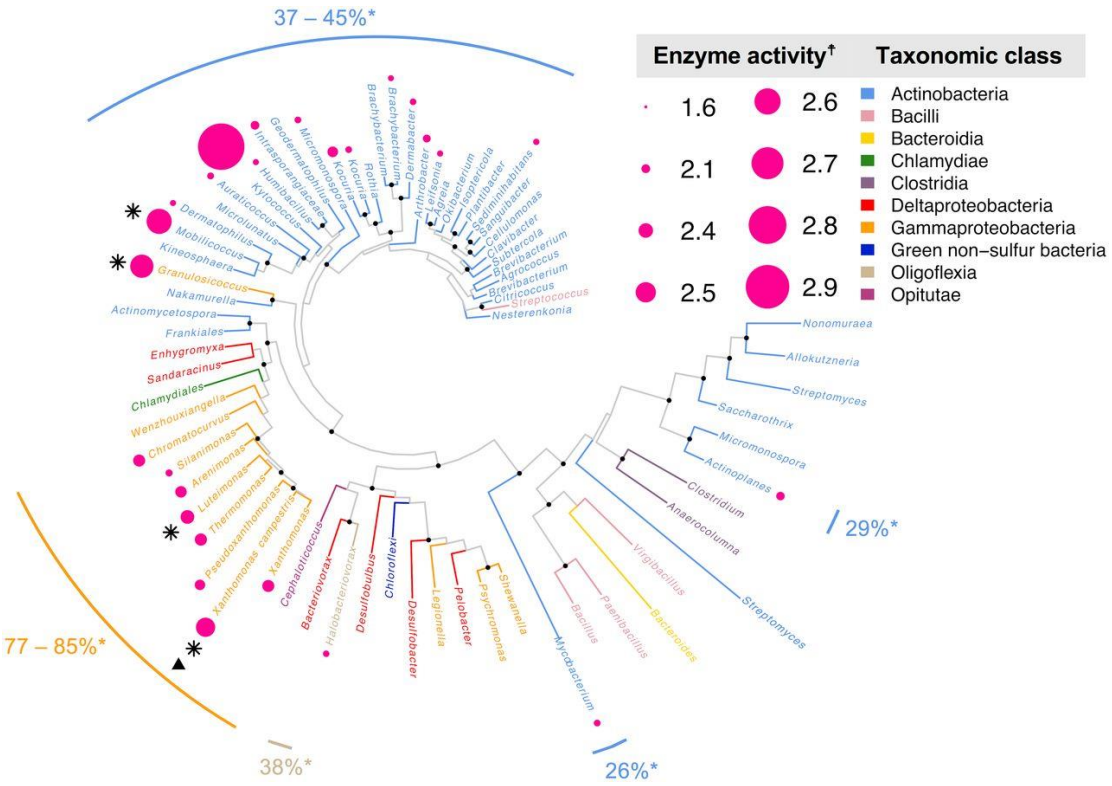


Figure 2-4. Phylogenetic tree of 73 taxonomically diverse OleA proteins assayed in this study using *p*-nitrophenyl hexanoate. Enzyme activity is shown as the $\log_{10}$ of nanomoles of pNP produced over the course of 1 h by an E. coli BL21 culture with an OD of 1.0 that is heterologously expressing OleA. Dark pink circles are scaled to relative enzyme activity levels measured in this study. Approximate maximum-likelihood phylogenetic analysis revealed three taxonomic classes of OleA homologs active with pNP esters: *Gammaproteobacteria* (orange), *Actinobacteria* (blue), and *Oligoflexus* (tan).

Four of these proteins were purified to homogeneity and assayed with various chain length acyl-CoA substrates (black stars, see Figure 2-5). Percentages marked with asterisks correspond to amino acid (aa) identity relative to *X. campestris* OleA (black triangle). Some actinobacterial sequences with as low as 26% aa identity to *X. campestris* are active with pNP esters. Branches are shown in color by the taxonomic class of the source organism. Black nodes correspond to branch points with probabilities of > 0.75.

There were clearly two major taxonomic clades that yielded active OleA proteins at a higher frequency than the cumulative 35% success rate. Indeed, the gammaproteobacterial proteins that were most similar (77 to 88% identity [ID]) to the *X. campestris* OleA were almost uniformly reactive with *p*-nitrophenyl hexanoate, giving an 88% success rate. However, in that group, none showed a higher activity than the characterized *X. campestris* OleA. Interestingly, a cluster of actinobacterial OleA proteins, and a related gammaproteobacterial sequence from *Granulosicoccus*, showed significantly higher activity than any of the proteins clustering with the *X. campestris* OleA (Figure 2-4). The highest activity overall was observed with an OleA homolog from *Kytococcus sedentarius*, a bacterium isolated from a marine environment in 1944, but also commonly found on human skin (57). The second highest activity was observed with an OleA from *Mobilicoccus massiliensis*, a bacterium isolated from a human stool sample (58). The *M. massiliensis* OleA amino acid sequence is only 37% identical to that from *X. campestris*, and the *Kytococcus* OleA amino acid sequence is likewise only 45% identical to the *X. campestris* OleA. *Kytococcus* and *Mobilicoccus* are both members of the order *Micrococcales*. The taxonomic outlier represented in this cluster of OleA sequence space is *Granulosicoccus antarcticus*, which is a marine gammaproteobacterium (59). The sequence of the *Granulosicoccus* OleA is fairly divergent, showing only 40 to 51% amino acid sequence identity to the other highly active proteins from *Xanthomonas*, *Mobilicoccus*, and *Kytococcus*. Several other actinobacterial OleA proteins showed activity, including proteins from *Actinoplanes atraurantiacus*, and *Mycobacterium obuense*. A.

atraurantiacus is a bacterium isolated from forest soil (60), and from a completely different family of *Actinobacteria*, the *Micromonosporaceae*, that are known for their prolific production of secondary metabolites (61). Mycobacterium obuense is a nonpathogenic member of the *Mycobacterium* genus that is studied for potential use in bioremediation (62). Another active OleA is from Halobacteriovorax marinus, belonging to the *Oligoflexia* class, found in estuaries and known to prey on Gram-negative bacteria (63). The *Actinoplanes* and *Mycobacterium* proteins are the most divergent active proteins from the *X. campestris* OleA, showing only 29% and 26% amino acid sequence identity, respectively. In light of these large sequence and taxonomic differences, we chose selected proteins to confirm that they showed Claisen condensation activity with long-chain acyl-CoA substrates, characteristic of OleA proteins, and to further explore gene cluster differences, both of which provide insights into the biological function of these OleA proteins.

Purification of new OleA proteins and investigating Claisen reactivity. Reaction of the *p*-nitrophenyl acyl substrates with OleA proteins leads to two products, a fatty acid and *p*-nitrophenol. We have not detected evidence of Claisen condensation between acyl chains of *p*-nitrophenyl esters as observed with native acyl-CoA substrates. However, purified *X. campestris* OleA mutants in which the active site cysteine is mutated to a serine or alanine do not have *p*-nitrophenyl ester hydrolysis activity (Supplemental Figure 2-S3). This indicated that the acyl chain is undergoing a transesterification from the activated *p*-nitrophenyl ester to the active site cysteine, comparable to the transesterification from acyl coenzyme A to the cysteine in the physiological reaction. Analysis of the reaction of OleA with *p*-nitrophenyl esters of various chain lengths using gas chromatography (GC)-mass spectrometry (MS) did not produce any detectable β-keto acid product. Presumably, hydrolysis of the enzyme intermediate outcompetes the binding and subsequent Claisen condensation with a second acyl substrate. In this context, we sought to determine whether the newly

identified proteins would indeed catalyze a Claisen condensation reaction with acyl-CoA substrates.

To test this, several proteins were purified and assayed via the standard assay procedure (Figure 2-5). We selected recombinant E. coli clones expressing highly active OleA proteins from the genera *Granulosicoccus*, *Mobilicoccus*, and *Luteimonas*. We also chose the *Actinoplanes* OleA protein, as it was one of the most divergent with respect to sequence and the taxonomy of the native organism. The first three proteins were stable to purification via nickel affinity column chromatography. However, the *Actinoplanes* enzyme was produced in low yield and readily precipitated upon concentration. The proteins expressed well in and showed expected bands of ~37-kDa subunit molecular weights when analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (Figure 2-5).

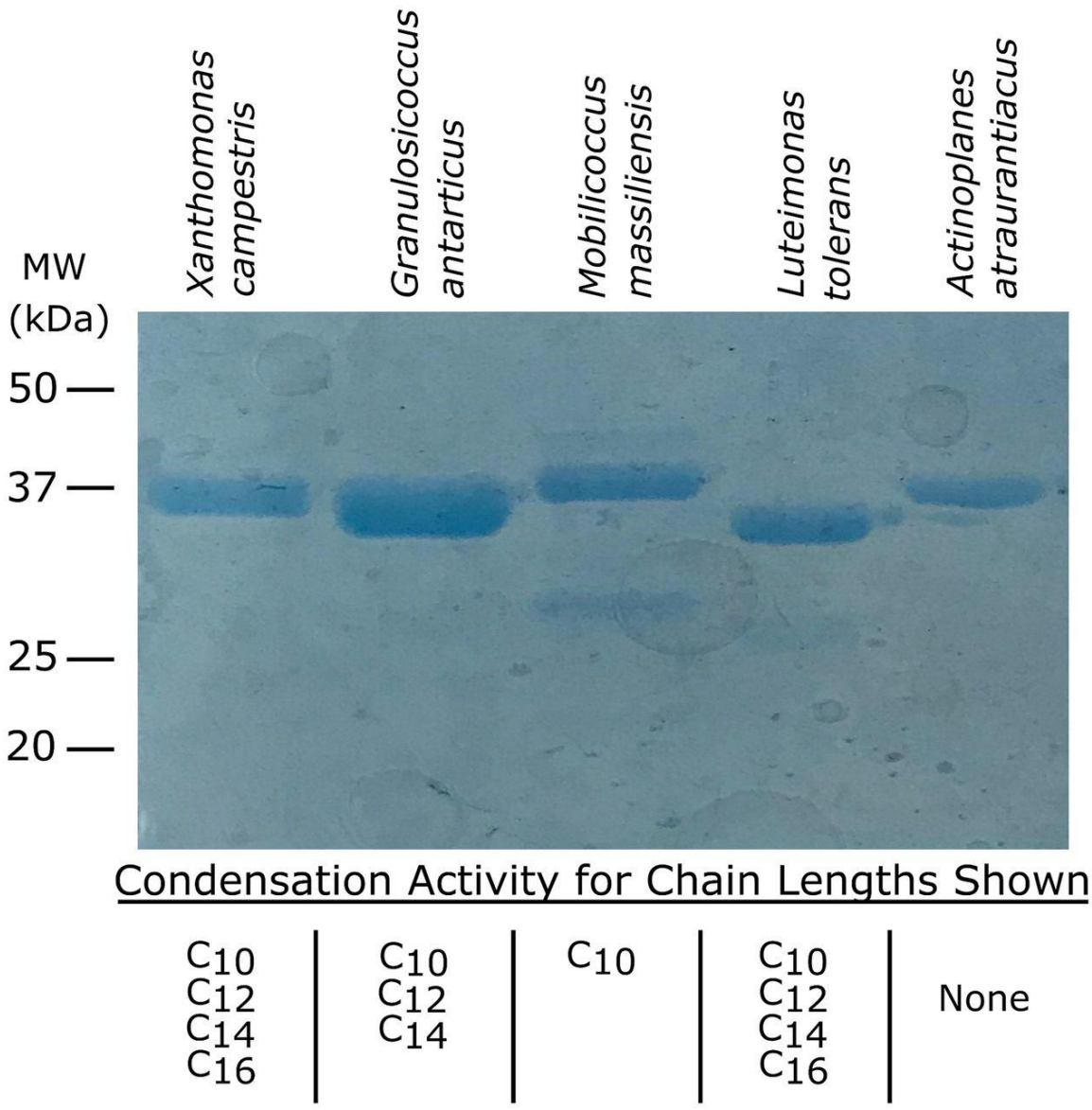


Figure 2-5 Purification, SDS-PAGE analysis, and assay of OleA proteins for Claisen condensation activity with long-chain acyl-CoA substrates. Each protein was assayed for Claisen condensation reactivity with acyl-CoA substrates ranging from C_8 to C_{16} as described in Materials and Methods. The chains lengths that reacted are indicated at the bottom. MW, molecular weight.

Enzyme assays were run via the standard assay involving incubation, extraction, and GC-MS analysis to detect and confirm the expected Claisen condensation product (Figure 2-5). *X. campestris* OleA served as a positive

control, and negative controls consisted of blanks without enzyme. The positive control condensed acyl-CoA with chain lengths of C_{10}, C_{12}, C_{14}, and C_{16}, similar to previous experiments (29). The *Granulosicoccus* OleA condensed C_{10}, C_{12}, and C_{14} chains, the *Mobilicoccus* Ole condensed C_{10} chains, and the *Luteimonas* OleA condensed C_{10}, C_{12}, C_{14}, and C_{16} chains, similar to the *Xanthomonas* enzyme. Limited *Actinoplanes* enzyme was available due to precipitation during purification, and the recoverable protein did not react with any of the acyl-CoA substrates tested.

Gene context and biological function of OleA proteins. The biological products from OleA-initiated metabolic pathways may be inferred from their genomic context and the fatty acyl chain pool of the native, producing organism (Figure 2-6). *X. campestris* condenses a number of saturated and unsaturated fatty acids to make long-chain olefinic hydrocarbons, shown previously by GC-MS analysis of membrane extracts (36). The *X. campestris* genome region contains two additional genes encoding Pfam domains PF13784 and PF03692, respectively, in addition to the *oleABCD* genes encoding proteins with known function. The related gammaproteobacterial protein from *Luteimonas* has a similar *ole* gene region but is lacking the PF13784 domain. The significance of these domains is currently unknown. On the basis of previous work showing that major fatty acid types are condensed to form olefins (15), we predict that branched-chain hydrocarbons are formed by *Luteimonas* (Figure 2-6). *Mobilicoccus* and *Granulosicoccus* have similar gene regions, but somewhat different membrane hydrocarbons are predicted based on differences in fatty acid content. A major feature of both of those gene regions is the presence of an *oleBC* gene fusion that tethers the OleB and OleC enzyme activities together into one polypeptide (6). This has been observed previously to occur in certain *Actinobacteria* typically in the order *Micrococcales*. The presence of the *oleBC* gene fusion in *Granulosicoccus* is the first observation of this outside the *Actinobacteria*, and the similar gene architecture may in fact be reflective of a horizontal gene transfer event.

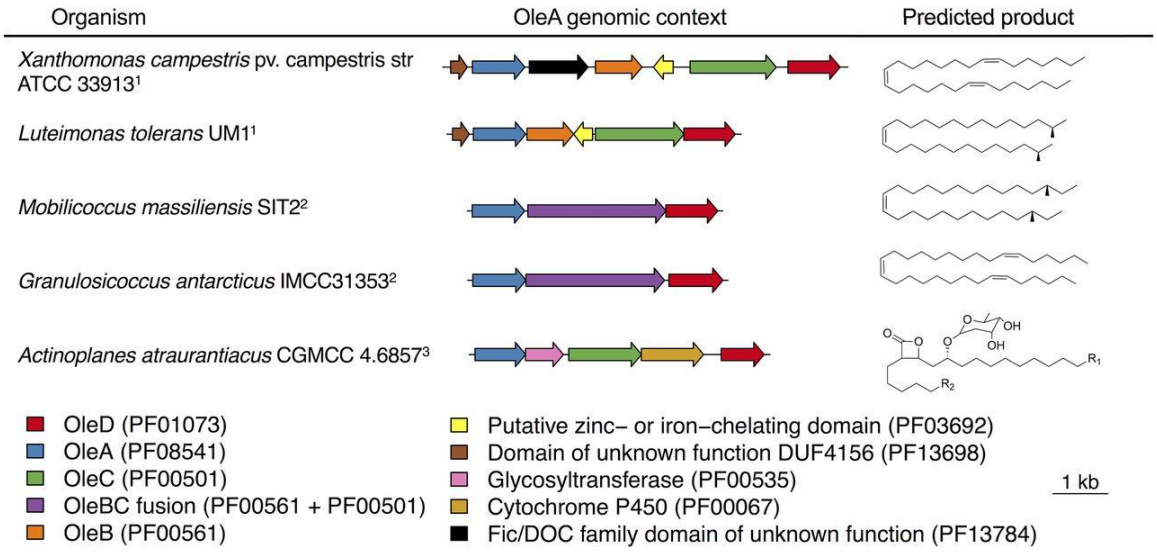


Figure 2-6. Genomic context for genes encoding OleA proteins that were expressed and purified in this study. Coloring corresponds to PFAM domain(s) present in each gene cluster. Analysis revealed three types of genome contexts: [1]*oleABCD* clusters with 0 to 3 intermediate genes, [2]*oleBC* fusion genes in *oleABCD* clusters, and [3]*oleACD* clusters lacking *oleB* homologs likely making β-lactone natural products (e.g., lipstatin analogs). Predicted products are inferred based on biosynthetic gene cluster synteny with known natural products, the predominant fatty acid composition of source organisms, and functional knowledge of the protein domains flanking OleA.

The *A. atraurantiacus* gene region is the most different, and we believe that it reflects a different biological function for the *ole* genes. The entire *A. atraurantiacus* genome lacks a candidate *oleB* gene. OleB makes a β-lactone decarboxylase enzyme (27) that acts in the membrane biosynthesis pathway to transform β-lactones into an olefinic hydrocarbon (Figure 2-1). The absence of this gene infers that the final product is a β-lactone. Indeed, other *Actinobacteria*, such as *Streptomyces toxytricini* and *Nocardia brasiliensis* that have *oleACD* biosynthetic gene clusters but lack an *oleB* gene produce β-lactone natural products (12, 41). The annotated gene region from the *A. atraurantiacus* genome indicates the following protein domains: OleA, a glycosyltransferase, OleC, a

cytochrome P450, and an OleD. That cluster of protein domains suggests a Claisen condensation of fatty acyl groups (OleA), reduction of the condensed product (OleD), hydroxylation of one of the acyl chains (cytochrome P450), glycosylation of the alcohol (transferase), and ring closure to make the β-lactone (OleC). The exact positions of the substituents cannot be determined, but the structure shown in Figure 2-6 is representative of the type of product that could be made by *A. atraurantiacus.*

Discussion

An impediment to more broadly studying diverse OleA proteins has been their poor expression/activity in *E. coli* and the slow published assay that required fixed time point sampling, extraction, and chromatography. This was overcome in the present study by the discovery and standardization of a rapid, sensitive, color-based assay. Many proteins have been shown to catalyze hydrolysis of *p*-nitrophenyl esters (49–53, 64), but to our knowledge, the Ole proteins are the first members of the thiolase superfamily to perform this reaction. For a control, we tested here the thiolase enzymes Pks13 from *Mycobacterium tuberculosis* and FabH from *E. coli* with *p*-nitrophenyl hexanoate and found no discernible reactivity. This helps explain the low background activity of *E. coli* enabling the development of a whole-cell assay. Most reported enzymes reacting with *p*-nitrophenyl esters are serine hydrolases such as lipases and proteases. Porcine intestinal lipase was tested here and showed only 13% of the specific activity observed with OleA when assayed with *p*-nitrophenyl hexanoate (see Supplemental Table 2-S2). Many lipases and other reactive serine enzymes are assayed with *p*-nitrophenyl acetate, which is much more soluble than its longer-chain counterparts. However, this substrate was not suitable for the *in vivo* assay developed here, as *p*-nitrophenyl acetate gave a very high background rate of hydrolysis in wild-type *E. coli* cells. In order to mitigate against false-positive results, we included only proteins showing a significant, reproducible level of activity, as discussed in Materials and Methods. The strict criteria used here eliminated consideration of a significant number of OleA proteins with low

activity. Indeed, based on the heatmap shown in Supplemental Figure 2-S2 in the supplemental material, more than two thirds of the OleA enzymes tested showed some measurable level of activity above background. Given that OleA proteins in divergent bacteria produce different products (36) (Figure 2-6), it is not surprising that some enzymes would not optimally bind and react with *p*-nitrophenyl hexanoate. Using positive results from the screen that fell within the cutoff, we had a 75% success rate of purification of our four chosen enzymes, which is much higher than the 20% success rate for five candidate enzymes examined in previous research (29).

The high rate of OleA in hydrolyzing *p*-nitrophenyl hexanoate versus other E. coli proteins allowed for a rapid assay to screen a wide range of putative OleA proteins that could be identified by bioinformatics. The only well-characterized OleA thus far has been the protein from *X. campestris*, a homodimer in which a glutamate from one subunit acts as a general base in the active site of the second subunit. This mechanism is unique to OleA homologs compared to other characterized proteins in the thiolase superfamily. A multiple-sequence alignment of the OleA proteins showed that all of the proteins except two showed a glutamate at position 117 followed by a proline (Supplemental Figure 2-S5). This glutamate was found to be essential to the Claisen condensation in *X. campestris* OleA. Bioinformatic analysis, both published (36) and conducted here, indicate that there are more than one thousand such OleA-type proteins in GenBank likely involved in natural product or membrane biosynthesis. The number and diversity of operons harboring *oleA* genes across different phyla is expansive, suggesting that the substrate specificity of these proteins and the final pathway products extend well beyond what has been characterized thus far. Indeed, even with our strict cutoff, we saw activity with OleA proteins from *A. atraurantiacus* and *M. obuense.* These proteins are 29% and 26% identical to OleA from X. campestris. For a comparison, FabH from *E. coli* has a 27% sequence identity to *X. campestris* OleA yet showed no activity with *p*-nitrophenyl hexanoate. Additionally, although we were able to purify *A. actinoplanes* OleA, it quickly precipitated out of solution and was difficult to work with. These two pieces of

data highlight two major points of the assay. One is that we can use the first transesterification step to test for activity for OleA enzymes from even a wide range of diversity. Second, this general method may be used to study OleA activity with enzymes that may not be amenable to purification. This opens up a much greater diversity of OleA enzymes to study.

OleA enzymes with different substrate selectivity can be combined with broad-specificity OleD and OleC enzymes to generate diverse β-lactones, which are of interest for their medicinal properties. In this context, it would be very beneficial if *p*-nitrophenyl esters could substitute for acyl-CoA compounds in biocatalytic cascades for making β-lactones. *p*-Nitrophenyl esters are commercially available, relatively inexpensive, and easy to synthesize compared to their CoA counterparts. However, we have not observed Claisen condensation to occur with *p*-nitrophenyl alkanoates and the OleA enzymes. The reason is not immediately clear, since there is evidence that the reaction initiates similarly to the acyl-CoA reaction with transfer of the acyl group to the active site cysteine. When C-143 is mutated to an alanine or when OleA is incubated with cerulenin, an inhibitor shown to interact with the active site cysteine (32), *p*-NP hydrolysis does not occur above background levels (Supplemental Figure 2-S4). This is similar to the change in rate of CoA hydrolysis from acyl-CoA substrates (31). Studies investigating the potential OleA-catalyzed Claisen condensation of *p*-nitrophenyl alkanoates are under way.

Overall, this new assay allows for rapid analysis of libraries of diverse OleA homologs. This has implications not only for finding other novel β-lactones in nature but also for finding novel bioproducts such as surfactants for which pathways are initiated by OleA. Additionally, this provides a solid basis for understanding OleA proteins for enzyme engineering to broaden substrate specificity. While this study utilized *p*-nitrophenyl hexanoate, one could imagine screening a similar library using *p*-NP esters that are very different from the natural alkyl chain substrates, leading to the biosynthesis of novel β-lactones.

Materials and methods

Chemicals and reagents. The following *p*-nitrophenyl acyl esters were obtained from Sigma-Aldrich (St. Louis, MO): acetate, butyrate, octanoate, decanoate, myristate, and palmitate. *p*-Nitrophenyl hexanoate was obtained from Tokyo Chemical Industry. Coenzyme A, Tris-HCl, *p*-nitrophenol, lysogeny broth mix, granulated agar, and methyl tertiary-butyl ether were also obtained from Sigma-Aldrich. Polymyxin B sulfate was obtained from Alfa Aesar. Isopropyl-β-d-1-thiogalactopyranoside (IPTG) and kanamycin were obtained from GoldBio.

Computational methods and phylogenetics. Seed sequences were selected from 16 known and highly likely OleA enzymes from genes in organisms producing long-chain olefin products. The sequences were structurally aligned using T-Coffee Expresso (65) and used to build a profile hidden Markov model (pHMM) specific for OleA enzymes using HMMER3 (66). The pHMM was searched against a custom database of 47,093 nonredundant RefSeq protein sequences containing at least one 3-oxoacyl-acyl carrier protein synthase III domain (PF08541). The pHMM hits were trimmed to a stringent E-value cutoff of $1e^{-42}$ to yield 920 unique OleA-like protein sequences. The flanking genes within a six-gene window on either side of each *oleA* homolog were pulled using RODEO (67). There were 251 *oleA* homologs with at least two flanking pathway genes (*oleB*, *oleC*, or *oleD*) within the same gene neighborhood as determined by pHMM. These 251 “high-confidence” *oleA* sequences with flanking hydrocarbon or β-lactone biosynthetic genes were further filtered to include only sequences with a length of 500 amino acids or less. The remaining 234 sequences were then clustered using CD-Hit at 50% sequence identity cutoff with a word size of 2 to obtain 41 cluster representatives. Of these, 27 *oleA* genes were selected on the basis of taxonomic diversity. The remaining 46 *oleA* genes were manually selected. To conduct phylogenetic analysis, amino acid sequences for 234 candidate OleA proteins were aligned using DECIPHER (68). The alignment was trimmed, and FastTree was used with default parameters to infer the approximate maximum-likelihood phylogeny using the Jones-Taylor Thornton model with CAT approximation (69).

Bacterial strains and growth conditions. E. coli T7 strains (catalog no. C2566I; New England BioLabs [NEB]) were used that contained a pET-28b+ vector. One strain was a vector-only control, and another contained *oleA* from *X. campestris* (NP_635607.1). Homologs of OleA found in other strains were synthesized, placed into a pET-28b+ vector, and transformed into T7 Express Competent E. coli by the Joint Genome Institute. Cells were grown at 37°C in lysogeny broth medium until an optical density (OD) of 0.3. They were then induced with 1 mM IPTG and incubated at 16°C overnight. Experiments using these cells follow the protocols below.

***In vitro* plate assay for OleA.** Tris-HCl (50 mM) (pH 8.0) was added to a 96-well flat-bottom suspension culture plate (catalog no. 25-104; Genesee Scientific) containing 4 µg of OleA, 5% ethanol, and 200 µM *p*-NP alkanoate in a total volume of 200 µl. Absorbance was read at 410 nm every minute for 30 min at 37°C. A standard curve was developed in a buffer containing 50 mM Tris-HCl at pH 8.0. To precisely determine the extinction coefficient for *p*-nitrophenol under the conditions used here, commercially available *p*-nitrophenol was added in different concentrations to 50 mM Tris-HCl (pH 8.0) and 5% ethanol. Absorbance was read at 410 nm in cuvettes with a 1-cm path length in a SpectraMax Plus 384 microplate reader (Molecular Devices). The path length for the 200-µl reaction mixture in 96-well plates was determined by adding known concentrations of the same commercially available *p*-nitrophenol to the 50 mM Tris-HCl (pH 8.0) buffer with 5% ethanol and using the calculated extinction coefficient and absorbance read at 410 nm to determine the path length to be 0.58 cm. The extinction coefficient under these conditions was determined to be 15,546 $M^{-1}\,cm^{-1}$, and this value was used for subsequent calculations. All reactions were run in parallel on the same plates as three to five replicates for data determination. Controls of each *p*-nitrophenyl alkanoate chain length containing no protein were used to normalize against any nonenzymatic hydrolysis in buffer.

***In vivo* plate assay for OleA.** E. coli BL21(DE3) cells were transformed with either empty pET-28b+ vector or the vector containing an *oleA* gene. Cells were induced with 100 µM isopropyl-β-d-thiogalactopyranoside when the absorbance reached 0.3 (600 nm). Induced cells were grown overnight at 16°C followed by resuspension in 50 mM Tris-HCl (pH 8.0) buffer at 0.1 OD. Resuspended cells were then diluted 16-fold and added to a 96-well microtiter plate containing 5% ethanol and 200 µM *p*-nitrophenyl alkanoate for a total volume of 200 µl. Absorbance was read every minute for 60 min at 410 nm at 37°C. Absorbance was normalized to induced cells containing empty pET-28b+ vector to account for any hydrolysis of the respective *p*-nitrophenyl alkanoate in buffer plus hydrolysis by any other E. coli enzymes other than OleA.

The use of cell lysis reagents and solubilizers of long-chain *p*-nitrophenyl esters was also tested to determine the effects on the observed rates. In those experiments, 63 µM polymyxin B sulfate and 10 mM methyl-β-cyclodextrin were added to cell suspension in the wells and allowed to incubate at room temperature for 2 h before the addition of *p*-NP.

Data analysis. Absorbance values were normalized by subtracting the absorbance values for cells with an empty pET-28b+ vector control. Absorbance was converted to nanomoles of *p*-NP produced by an E. coli BL21 culture with an OD of 1.0 using the Beer-Lambert law (ε_{410} = 15,546 M^{-1} cm^{-1}). Slopes were calculated using a rolling linear regression window method by calculating slopes for all overlapping 15-min intervals over the course of the first 45 min of each reaction. The greatest slope for each enzyme with $R^2 \geq 0.9$ was selected as maximum enzyme activity for a given OleA (nanomoles of *p*-NP/OD of 1.0/hour). Enzyme activity values across triplicate measurements were averaged. Activity values displayed a right skewed distribution; therefore, a $\log_{10}$ transformation was applied for downstream analysis, resulting in an approximately normal distribution. Outliers more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile were removed to calculate mean enzyme activity. To set a stringent threshold for activity and filter out false-positive results,

we labeled enzymes that were greater than or equal to the mean enzyme activity of 1.66 $\log_{10}$ nmol *p*-NP/OD of 1.0/hour) as “active,” while those below the mean level of activity were considered “inactive.”

Protein purification and characterization. OleA proteins were purified, and assays with acyl-CoA compounds were performed as previously described for *X*, *campestris* OleA (29).

Acknowledgements

We thank Kelly Aukema for helpful discussions and a critical reading of the manuscript.

M.D.S. acknowledges support from a National Institutes of Health Biotechnology training grant (5T32GM008347-27). S.L.R. acknowledges support from a National Science Foundation Graduate Research Fellowship (grant 00039202). We thank the U.S. DOE for providing synthetic DNA to L.P.W. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by contract DE-AC02-05CH11231.

CHAPTER 3 – Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily

---

"Alone we can do so little; together we can do so much" – Hellen Keller

Synthetic Biology, 2020, 5(1), ysaa004
DOI: 10.1093/synbio/ysaa004

Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily

Serina L Robinson, Megan D Smith, Jack E Richman, Kelly G Aukema, Lawrence P Wackett

Summary

Enzymes in the thiolase superfamily catalyze carbon–carbon bond formation for the biosynthesis of polyhydroxyalkanoate storage molecules, membrane lipids and bioactive secondary metabolites. Natural and engineered thiolases have applications in synthetic biology for the production of high-value compounds, including personal care products and therapeutics. A fundamental understanding of thiolase substrate specificity is lacking, particularly within the OleA protein family. The ability to predict substrates from sequence would advance (meta)genome mining efforts to identify active thiolases for the production of desired metabolites. To gain a deeper understanding of substrate scope within the OleA family, we measured the activity of 73 diverse bacterial

thiolases with a library of 15 *p*-nitrophenyl ester substrates to build a training set of 1095 unique enzyme–substrate pairs. We then used machine learning to predict thiolase substrate specificity from physicochemical and structural features. The area under the receiver operating characteristic curve was 0.89 for random forest classification of enzyme activity, and our regression model had a test set root mean square error of 0.22 (R^2 = 0.75) to quantitatively predict enzyme activity levels. Substrate aromaticity, oxygen content and molecular connectivity were the strongest predictors of enzyme–substrate pairing. Key amino acid residues A173, I284, V287, T292 and I316 in the *Xanthomonas campestris* OleA crystal structure lining the substrate binding pockets were important for thiolase substrate specificity and are attractive targets for future protein engineering studies. The predictive framework described here is generalizable and demonstrates how machine learning can be used to quantitatively understand and predict enzyme substrate specificity.

Introduction

Metabolic pathways for the β-oxidation of fatty acids and production of polyketides, surfactants, β-lactone natural products and hydrocarbons are initiated by enzymes in the thiolase superfamily (6, 36, 47, 70). Carbon–carbon bond formation from the Claisen condensation of two activated fatty-acyl substrates by enzymes in the OleA family of thiolases (29, 36) represents the first committed step in production of the backbone of many value-added bacterial metabolites (Figure 3-1A). Previously, we demonstrated that swapping OleA from *Shewanella oneidensis* with *Stenotrophomonas maltophilia* OleA altered the chain length and profile of hydrocarbons produced downstream by OleBCD enzymes (36). Within the broader thiolase superfamily, Prather and colleagues used a rational design approach to alter thiolase substrate specificity for metabolic engineering of the reverse β-oxidation pathway (48). Thiolase substrate specificity is particularly critical for metabolic engineering applications since it often ‘sets’ the chemical composition of downstream products.

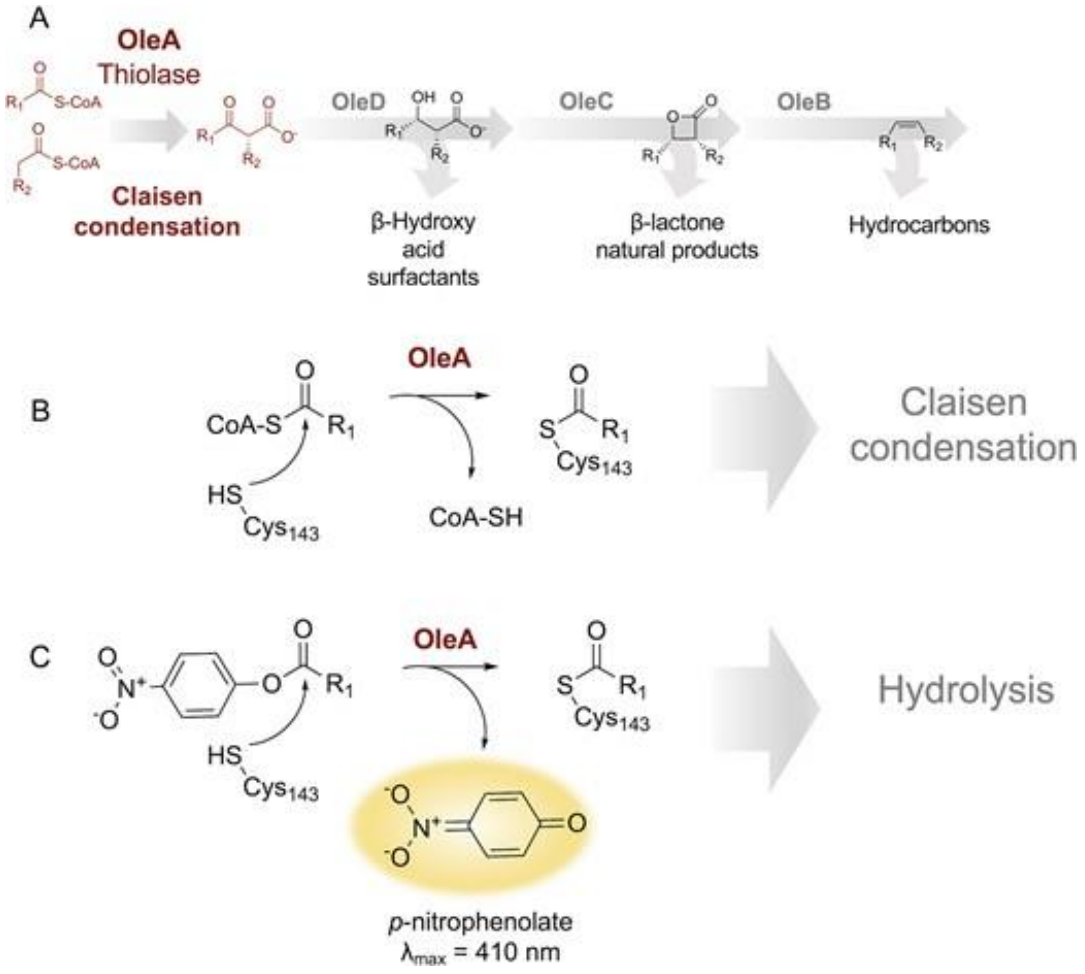


Figure 3-1. (A) Thiolase enzymes in the OleA family catalyze head-to-head Claisen condensation of two acyl-CoA substrates (maroon) as the first committed step in production of value-added metabolites such as surfactants, pharmaceuticals and hydrocarbons. R_1, R_2 in characterized OleABCD pathways (3): C_8 –C_{16} (B) Acyl-CoAs as native substrates for OleA-type thiolases acylate the active site cysteine (Cys 143 in *X. campestris*) prior to carbon–carbon bond formation via a Claisen condensation and substrate release. (C) OleA reacts with various *p*-nitrophenyl esters as substrate mimics to produce a *p*-nitrophenolate chromophore that absorbs in the visible range at pH 8.0, providing a rapid readout for enzyme activity.

Thiolases typically use a ping-pong mechanism whereby an activated substrate acylates the active site cysteine and remains tethered covalently until the second substrate binds and the acyl group is transferred, resulting in carbon–carbon bond formation. The majority of well-characterized thiolases are FabH-type enzymes with a single, deep hydrophobic substrate channel (71). FabH initiates the two-carbon elongation of fatty acids iteratively by condensing a malonyl unit onto a growing backbone with the concurrent release of CO_2 during each catalytic cycle. In contrast, OleA-type thiolases have two hydrophobic substrate channels instead of one and catalyze the nondecarboxylative Claisen condensation of two long-chain fatty acids (31, 32). At present, the only biochemically and structurally characterized OleA is from *Xanthomonas campestris,* a bacterial plant pathogen (29, 31, 32).

OleA enzymes characterized to date accept fatty acid substrates activated with coenzyme A (CoA), which are costly feedstocks for biotechnological applications. During the first step of the OleA catalytic cycle, acyl-CoA substrates undergo transesterification to the active site cysteine (Figure 3-1B) prior to carbon–carbon bound formation (29, 31, 32). Recently, we discovered that OleA-family enzymes also hydrolyze *p*-nitrophenyl esters (*p*NPs) to release *p*-nitrophenolate as a rapid colorimetric readout for OleA activity (Figure 3-1C). Here, we used this assay to screen 15 different *p*NP substrates against a library of 73 OleA sequences from taxonomically diverse bacteria sharing as low as 13.8% pairwise amino acid identity. We then trained machine learning models on our paired enzyme–substrate dataset to quantitatively predict thiolase activity with different *p*NPs.

Machine learning is gaining traction in chemical biology as a powerful technique for the prediction of enzyme substrate specificities (72). Support vector machines and ensemble learning methods achieved high accuracy for the prediction of amino acid substrates for nonribosomal peptide synthetase adenylation domains (73, 74). Integration of these machine learning algorithms within a larger predictive pipeline known as antiSMASH has improved structural prediction of natural products from genomic information (75). Machine learning-

based methods have also been applied to predict substrate specificities of other protein families including glycosyltransferases (76), acyl-CoA ligases (39) and proteases (77–79).

While thiolases have many applications in synthetic biology, quantitative insights into how physicochemical properties of residues in the substrate binding pockets affect substrate specificities are lacking. Using thiolases as biological catalysts to produce desired compounds requires a deeper understanding of their natural substrate range. Accurate prediction of substrate specificity will aid in expanding the toolbox of standardized parts for carbon–carbon bond formation and enable enzymatic production of compounds with backbones of desired chain length and composition. The experimental and machine learning frameworks described here may also be generalized to learn the substrate specificity rules of different enzymes classes. This approach can be integrated into an automated workflow for machine learning-guided parts selection for custom metabolite production and other applications in synthetic biology.

Materials and methods

Chemicals and reagents. Chemical syntheses of nine *p*NPs from their corresponding carboxylic acid precursors was carried out using the method of Engström *et al.* (80). Detailed synthetic procedures are in the Supplemental Methods, and proton nuclear magnetic resonance (^{1}H-NMR) spectra of all synthesized compounds are in Supplemental Figure S1. Purchased carboxylic acids from Sigma-Aldrich were 3-cyclopentylpropanoic acid, 7-phenylheptanoic acid, 3-(4-chlorophenoxy)propanoic acid, 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoic acid, 2-(2-butoxyethoxy)acetic acid and 6-heptynoic acid. 2,2-Dimethylhexanoic acid and heptanoic acid were purchased from TCI-America, and 6-azidohexanoic acid was purchased from Cayman Chemical Company. Additionally, five commercially available *p*NPs were obtained from the companies indicated: *p*-nitrophenyl trimethylacetic acid, *p*-nitrophenyl decanoic acid and *p*-nitrophenyl dodecanoic acid (Sigma-Aldrich), *p*-nitrophenyl biotin and *p*-

nitrophenyl benzoic acid (Alfa Aesar) and *p*-nitrophenyl hexanoic acid (TCI-America).

Strain and plasmid construction. Synthetic DNA for a library of 73 OleA homologs were cloned with N-terminal 6×-His tags into pET28b+ vectors at NdeI and XhoI restriction sites and transformed into T7 Express Competent Escherichia coli cells (NEB C2566I) by the U.S. Department of Energy Joint Genome Institute. Accession numbers (Supplemental Table 3-S2) and codon-optimized plasmid sequences (Supplemental Material 3-S1) are available.at 10.5281/zenodo.3743415. Bioinformatic methods for gene selection and construct design are described in detail by Smith *et al.* (43).

Culture conditions and whole-cell assays. Cells were grown at 37°C shaking at 250 rpm in 5 ml of lysogeny broth in test tubes to an optical density of 0.3 and induced with a final concentration of 1 mM isopropyl β-D-1-thiogalactopyranoside. Induced cultures were incubated for 40–43 h at 16°C with agitation at 250 rpm. Substrate specificity screening was conducted using the whole-cell thiolase assay protocol described by Smith *et al.* (20). Briefly, cells were normalized to an optical density of 0.1 per 200 µL of cells in 50 mM Tris–HCl (pH 8.0) buffer and incubated for 2 h with 63 µM polymyxin B sulfate to render them porous to small molecules. Cells expressing each of the 73 enzymes were tested with 15 different *p*NP substrates added at a maximum final concentration of 200 µM. Poor solubilities of some substrates resulted in minor precipitation. In a previous study, it was demonstrated that adding substrates above their limit of solubility did not significantly interfere with assay readout (43). Absorbance was read in 96-well clear bottom plates (Genesee Scientific) every minute for 60 min at 410 nm and 37°C using a SpectraMax Plus 384 microplate reader (Molecular Devices). All 73 enzymes were tested with 15 *p*NP substrates, first with technical duplicates from the same induction batch, and then in biological triplicates with three independently grown cultures for all of the active enzyme–substrate pairs.

Data preprocessing. Absorbance values were normalized by subtracting averaged values from triplicate empty pET-28b+ vector controls on each plate. Absorbance values for each plate were converted to nmol *p*-nitrophenolate using *p*-nitrophenolate standard curves run in parallel on the same plate. Activity was calculated using a 'rolling window' linear regression method by calculating slopes for all overlapping 15 min intervals over the course of the first 45 min of each reaction, with the exception of *p*-nitrophenyl 2-(2-butoxyethoxy)acetate which hydrolyzed rapidly in buffer such that catalyzed reaction rates above background could only be measured for 5 min before absorbance reached the maximum detection limit. The maximum slope with an $R^2 \geq 0.9$ was converted to enzyme activity (nmol *p*-nitrophenolate/OD 1.0/h). Activities for each enzyme–substrate pair were averaged across biological triplicates and are reported in Supplemental Table S2.

Physicochemical feature engineering. For each of the 15 *p*NP substrates, 153 chemical properties were calculated using the *Rcpi* and *ChemmineR* packages in R (81, 82). Since many of the chemical properties were correlated, we performed dimensionality reduction using principal component analysis. Loadings of the first seven principal components included in analysis are detailed in Supplemental Figure S2. To extract protein sequence features, we used DECIPHER, a structure-based aligner (68) that uses local sequence context to align each of the 73 OleA sequences with the crystal structure of *X. campestris* (PDB ID: 4KU5). We extracted spheres of residues from each aligned protein with radii 8, 10, 12 and 14 Å from the α-carbon of the active site cysteine. Twelve angstrom was selected as the best sphere size for model training because it was the smallest radius that encompassed all residues lining both substrate binding pockets, thereby keeping the number of features relative to total training data points reasonable. Each amino acid in the 12 Å radius was encoded as a vector of principal components of its physicochemical properties as described by Atchley *et al.* (83). The codon diversity index calculated by Atchley *et al.* was not included

analysis, therefore each of the 84 amino acids within a 12 Å sphere were encoded by four indices corresponding to polarity, molecular volume, secondary structure and electrostatic charge. Global protein properties used as features including instability index, isoelectric point, molecular weight and Kyte-Doolittle hydrophobicity index (see Supplemental Table 3-S3B at 10.5281/zenodo.3743415 for full set of indices) were calculated using the *Peptides* package in R (84).

Machine learning. Protein and chemical features were concatenated into a single numeric vector describing each enzyme–substrate pair ($n = 1095$). Features with near-zero variance were removed using the *nearZeroVar* function from the *caret* package in R (85). The data were split randomly with stratified sampling by activity to achieve roughly equal proportions of active and inactive enzymes in 75% training and 25% testing sets. R version 3.6.1 and *caret* were used to evaluate all models (85). Grid search was used to tune model hyperparameters by 10-fold cross validation repeated in triplicate. For the random forest algorithm, model hyperparameters that were tuned included the number of variables randomly sampled as candidates at each split, the minimum size of terminal nodes and the splitting rule methods. All forests were grown to a size of 1000 trees and the permutation method was used to compute relative feature importance as implemented in the *ranger* package in R (86). A description of hyperparameters tuned for other machine learning models tested were detailed in Supplemental Figure 3-S3. The distribution of training and testing set prediction performances were examined by 1000 independent, random training-test splits (Supplemental Figure 3-S4).

Three different machine learning algorithms were evaluated for classification of enzyme–substrate pairs: random forest, naïve Bayes and feedforward neural networks. Receiver operating characteristic curves and confusion matrices were used to compare model performances. For the 550 enzymes classified as active, we further trained regression models to quantitatively predict enzyme activity. Enzyme activity values displayed a right-

skewed distribution so a $\log_{10}$ transformation was applied, resulting in an approximately normal distribution. Three different machine learning algorithms were evaluated for regression: random forest, elastic net and multivariate adaptive regression splines. Root mean square error (RMSE) and R^2 values were used to assess performance. Models were further evaluated using leave-one-compound-out and leave-one-taxon-out validation (Supplemental Figure 3-S5).

Homology modeling, phylogenetics and bioinformatics. An approximate maximum-likelihood phylogeny for the OleA amino acid sequences was estimated using FastTree version 2.1 (69) using the Jones–Taylor-Thornton model and assuming a single rate of evolution for each site known as the "CAT" approximation. Between-groups analysis was used to detect differential residues between broad and narrow specificity enzymes using the *bgafun* package in R (87). Homology models for each of the 73 OleA proteins were built using the Phyre2 server (88). Solvent-accessible surface area and cavity volumes for each of the homology models were calculated using CastP version 3.0 with a 2.2 Å radius probe (89).

Data and code availability. Raw data and scripts to reproduce analyses, figures and tables are available at https://github.com/serina-robinson/thiolase-machine-learning/. An interactive web application with a searchable database and predictive models trained on the complete dataset are also available at z.umn.edu/thiolases (shortened URL) and srobinson.shinyapps.io/thiolases (permanent URL). The DNA constructs, provided in Supplemental Material 3-S1, will be provided upon request. The DNA constructs were provided by the United States Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231. DNA requests will be honored with the completion of a Materials Transfer Agreement as required by our contracts with the U.S. Department of Energy.

Results

Selection and screening of chemically diverse *p*NP substrates. We aimed to construct the first quantitative map of natural variation in substrate specificity for OleA-type thiolases. Recently, we developed a whole-cell assay using *p*NPs to rapidly screen thiolase enzymes without time-consuming protein purification, quenching and extraction steps (43). Here, we used this assay to screen a library of thiolase enzymes with 15 diverse *p*NP compounds. To select candidates for screening, we web scraped Sigma-Aldrich pages to identify commercially available carboxylic acids ($n = 3572$). Based on clustering analysis using the Tanimoto coefficient (Supplemental Figure 3-S6A), we selected 15 *p*NPs that spanned a wide range of chain lengths, heteroatomic composition and functional groups (Supplemental Figure 3-S6B). The final library of 15 *p*NPs were synthesized from their carboxylic acid precursors as described in the Supplemental Methods or purchased directly as described above.

We screened each of the 15 *p*NPs against a panel of 73 bacterial thiolase enzymes heterologously expressed in *E. coli* for a total of 1095 enzyme–substrate pairs (Figure 3-2). All enzymes were active with at least two *p*NP substrates and, with the exception of *p*-nitrophenyl benzoate, all substrates reacted with at least one enzyme. The most broadly reactive *p*NP was *p*-nitrophenyl 2-(2-butoxyethoxyacetate) for which 71 out of 73 enzymes yielded product. The enzyme with the highest average activity was natively from *Kytococcus sedentarius*, a common constituent of the human skin microbiome. The *K. sedentarius* thiolase reacted with 14 different *p*NPs and was among the top 3 most active enzymes for 10 of these substrates.

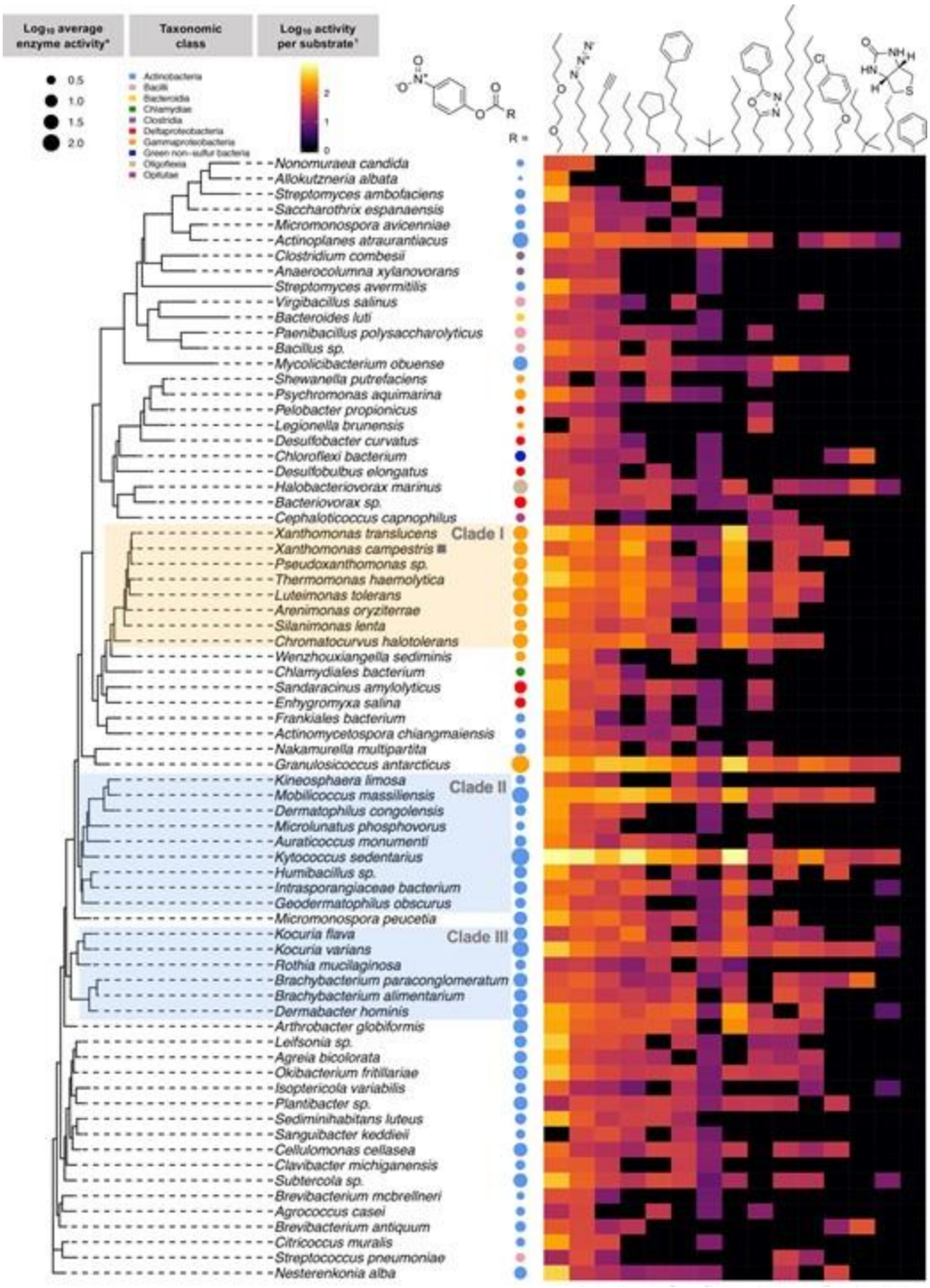


Figure 3-2. Approximate maximum-likelihood phylogeny of 73 OleA protein sequences paired with heatmap of enzyme activities across 15 different *p*NP substrates. †Enzyme activities were measured as the $\log_{10}$ of nmol *p*-nitrophenolate produced over the course of one hour by an *E. coli* BL21 culture

with an OD of 1.0 per 200 µL of cells heterologously expressing OleA. *Average enzyme activity is across three biological replicates for each substrate screened. Circle sizes are scaled to average activity for each enzyme across all substrates. Clade I (orange shading) represents the most active clade of Gammaproteobacterial enzymes including the *X. campestris* OleA designated with a gray square for which the crystal structure (PDB ID: 4KU5) has been solved (31, 32). Clades II and III (blue shading) represent the most active clades of OleA enzymes found in Actinobacteria.

Broad substrate specificity among actinobacterial and gammaproteobacterial OleA clades. We first examined whether OleA thiolase substrate specificity had a phylogenetic signal. Our enzyme library included sequences from 7 different phyla and 68 different bacterial genera (43). We observed three monophyletic clades that exhibited a high level of activity across a wide range of *p*NP substrates (Figure 3-2). Clade I contained gammaproteobacterial thiolases within the *Xanthomonadaceae* including *Chromatocurvus, Luteimonas, Thermomonas* and the structurally characterized *X. campestris* OleA (31, 32). The other two highly active clades (II and III) consisted of thiolases from Actinobacteria including *Kytococcus*, *Mobilicoccus, Dermatophilus* and *Kocuria*. Even within these clades, *p*NP substrate profiles were variable across enzymes from closely related organisms and activity could not be predicted on the basis of phylogeny alone.

Relationship between binding pocket volume and bulky substrate preferences. *Actinoplanes atraurantiacus* had the highest preference for *p*-nitrophenyl trimethylacetate across all enzymes in the library, even outperforming *K. sedentarius* and other highly active thiolases. *Brachybacterium paraconglomeratum* and *Halobacteriovorax marinus* also had thiolases with 'bulky' substrate specificity, exhibiting higher activity with *p*-nitrophenyl 2,2-dimethylhexanoate than with the C_6–C_7 length substrates preferred by the majority of enzymes screened. Since the *tert*-alkyl substituents of the

trimethylacetate and 2,2-dimethylhexanoate compounds fill a larger volume near the α-carbon than most of the other *p*NPs tested, we hypothesized that the volume of the substrate binding pocket might affect enzyme activity. We constructed homology models for the 73 enzymes and used a computational geometry method (89) to calculate the solvent-accessible surface area and binding pocket volumes (Figure 3-3). We observed a weak positive association between solvent-accessible surface area and average enzyme activity (Pearson correlation = 0.27, *P*-value = 0.02). Three enzymes with unusually high preferences for bulky α-carbon substrates (*Actinoplanes, Halobacteriovorax* and *Brachybacterium*) also had among the highest predicted solvent-accessible surface areas (Figure 3-3). However, we noted other enzymes with lower predicted solvent-accessible surface areas also displayed comparable activity with bulky substrates while some enzymes with large predicted surface areas did not display a preference for bulky substrates (Figure 3-3). Although care was taken to thread all models to the same template to avoid template bias, results must be interpreted with caution since homology modeling does not recapitulate structural dynamics and variations in pocket volume during substrate binding.

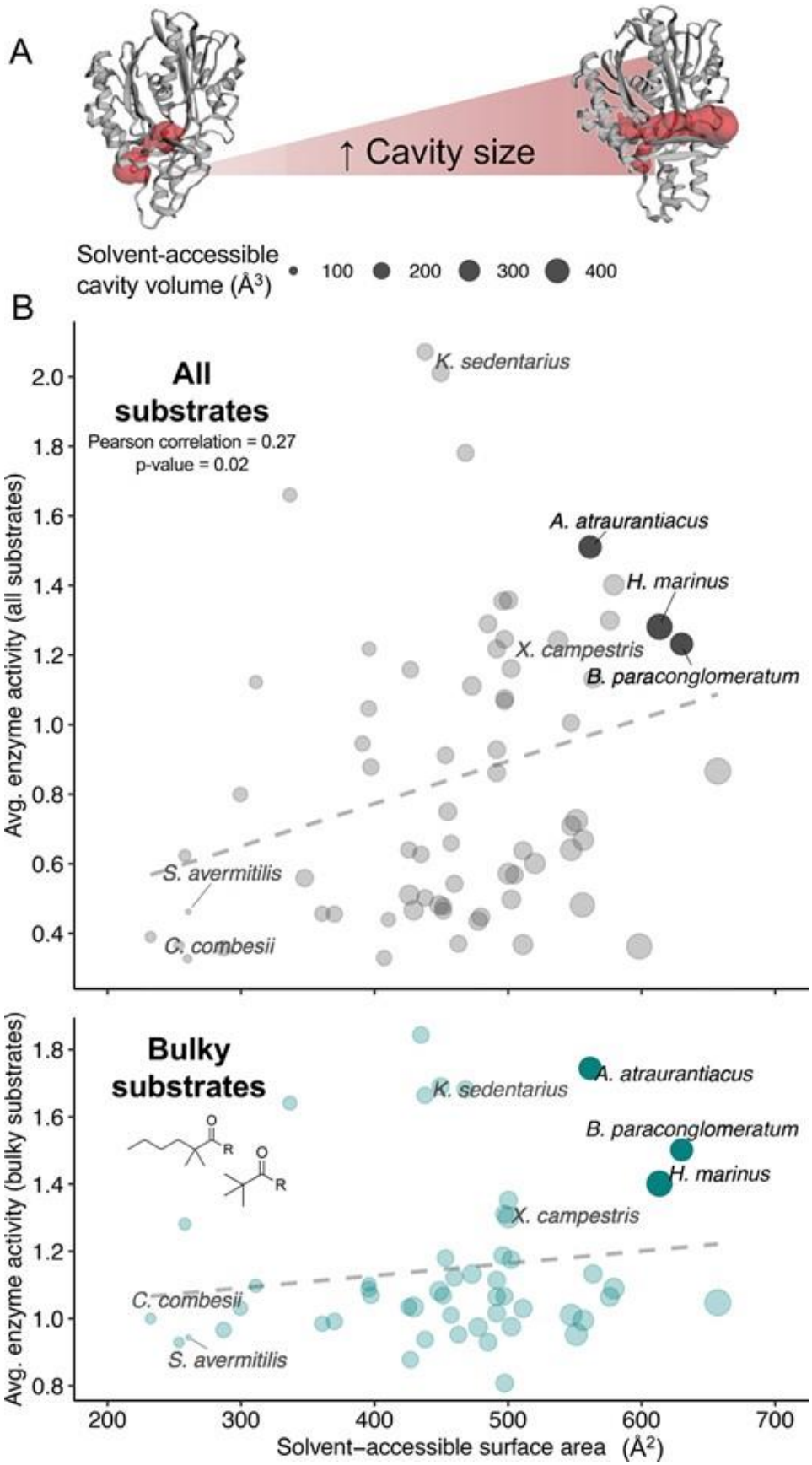

A
↑ Cavity size
Solvent-accessible cavity volume (Å³)
100
200
300
400
B
All substrates
Pearson correlation = 0.27
p-value = 0.02
K. sedentarius
A. atraurantiacus
H. marinus
X. campestris
B. paraconglomeratum
S. avermitilis
C. combesii
Avg. enzyme activity (all substrates)
Bulky substrates
Avg. enzyme activity (bulky substrates)
Solvent−accessible surface area (Å²)

Figure 3-3. (A) Predicted solvent-accessible cavity volumes of the 73 OleA enzymes analyzed in this study ranged from 95.3 to 478.9 $Å^3$. (B) Relationship between the solvent-accessible surface area with the average enzymatic activity across all substrates (black) and 'bulky' substrates with *tert*-alkyl-substituted α-carbons (teal, *p*-nitrophenyl trimethylacetate and *p*-nitrophenyl 2,2-dimethylhexanoate). Enzymes from *Actinoplanes atraurantiacus, Brachybacterium paraconglomeratum* and *Halobacteriovorax marinus* have among the highest calculated cavity volumes and the highest preferences for substrates with *tert*-alkyl-substituted α-carbons.

Statistical identification of specificity-determining residues T292 and L203. Approximately one-third of the enzymes we tested were active with 10 or more *p*NP substrates, suggesting that a large number of OleA-type thiolases had a broad substrate range, including the highly active enzyme from *K. sedentarius*. In contrast, enzymes that reacted with five substrates or fewer (Supplemental Table 3-S2, provided at 10.5281/zenodo.3743415) were defined to have narrow substrate specificity. We used a multivariate statistical method termed between-groups analysis (BGA) to detect residues that differed between broad- and narrow-specificity thiolases (87). BGA consistently identified two key residues lining substrate binding channels in the crystal structure that were conserved in more than 50% of enzymes in each group. The identification of these two residues was robust to different splits and group sizes for broad and narrow specificity enzymes. Residue 292 was located towards the end of the substrate channel A in the *X. campestris* crystal structure. It was a conserved Thr in broad specificity enzymes that was absent in enzymes with narrow specificity (*P*-value < 0.05, Fisher exact test). All but one of the top 30 most active enzymes in our dataset had a Thr aligning with position 292. Narrow specificity thiolases had variety of charged (Asp, Glu, Arg) or aliphatic (Ile, Leu, Val) residues aligning with position 292 instead. The second key residue aligned with residue 203 located at the end of substrate channel B in the *X. campestris* structure. This residue was a conserved Leu/Ile in the top 32 broad specificity enzymes and a

Gly or Val among narrow specificity enzymes that accepted four or fewer substrates (*P*-value < 0.05, Fisher exact test).

Machine learning prediction of paired enzyme–substrate activity relationships. While BGA was useful to identify conserved amino acids between enzyme groups that may affect substrate specificity, it could not fully capture the complexity of enzyme–substrate relationships; this task is better suited for machine learning. We next evaluated the performance of different machine learning algorithms to predict substrate specificity from a combination of physicochemical protein and substrate features.

To construct a set of substrate features, we calculated 153 chemical descriptors and used principal component analysis to reduce these descriptors into linearly uncorrelated principal components (PCs, Supplemental Figure 3-S2A). The first seven chemical PCs were able to explain 92% of the variance between substrates (Supplemental Figure 3-S2B). We extracted the absolute values of the PC loadings to determine the overall contribution of different chemical descriptors to the PCs (Supplemental Figure 3-S2C). We observed that the top seven PCs corresponded broadly to the following chemical properties: molecular weight (PC1), molecular connectivity (PC2), aromaticity (PC3), solubility (PC4), oxygen content (PC5), nitrogen content (PC6) and chlorine content (PC7).

A structure-based sequence alignment method was used to align each protein sequence in the training set with the crystal structure of an OleA-type thiolase with substrates bound (PDB ID: 4KU5). All residues that aligned within 12 Å of the active site cysteine (Cys 143) were encoded into a numerical vector of physicochemical indices corresponding to polarity, molecular volume, secondary structure and electrostatic charge. We found this method of amino acid featurization increased training set classification accuracy 2% over a one-hot encoding method which does not capture information about the physicochemical properties of amino acids. We also calculated macromolecular protein properties based on the full-length input sequences including

hydrophobicity, isoelectric point, molecular weight and instability indices (Supplemental Table 3-S3B, provided at 10.5281/zenodo.3743415). All chemical and protein features were then concatenated into a single vector representing the unique physicochemical signature of each enzyme–substrate pair.

Substrate aromaticity and molecular connectivity influences enzyme–substrate pairing. Of the 1095 enzyme–substrate pairs tested experimentally, 550 were active and 545 were inactive (Figure 3-4A). We first evaluated three different machine learning algorithms for binary classification of enzyme–substrate pairs as active or inactive. Of the three algorithms tested (random forest, naïve Bayes and feedforward neural networks), we observed the highest classification accuracy (area under the receiver operating characteristic curve = 0.89) with the random forest model (Table 3-1, Supplemental Figure 3-S3A). From 1000 random training-testing dataset splits we obtained an average testing set classification accuracy of 81.9 ± 2.2% (Supplemental Figure 3-S3A). Our model indicated that chemical features were universally more important than sequence features for classification of enzyme–substrate pairs as active or inactive (Figure 3-4B). In particular, the substrate aromaticity index (PC3) was the most important feature followed by the molecular connectivity index (PC2) including valence chi chain descriptors, Kier molecular shape indices and number of rotatable bonds (Supplemental Figure 3-S2, Supplemental Table 3-S3 provided at 10.5281/zenodo.3743415). Activity increased with carbon chain length to a certain point (7–8 carbons) beyond which activity decreased, as observed with *p*-nitrophenyl decanoate and *p*-nitrophenyl dodecanoate. Overall, the majority of enzymes tested preferred substrates with a higher number of rotatable C–C or C–O bonds and a chain length of 6–7 carbons instead of longer or shorter chain lengths tested here.

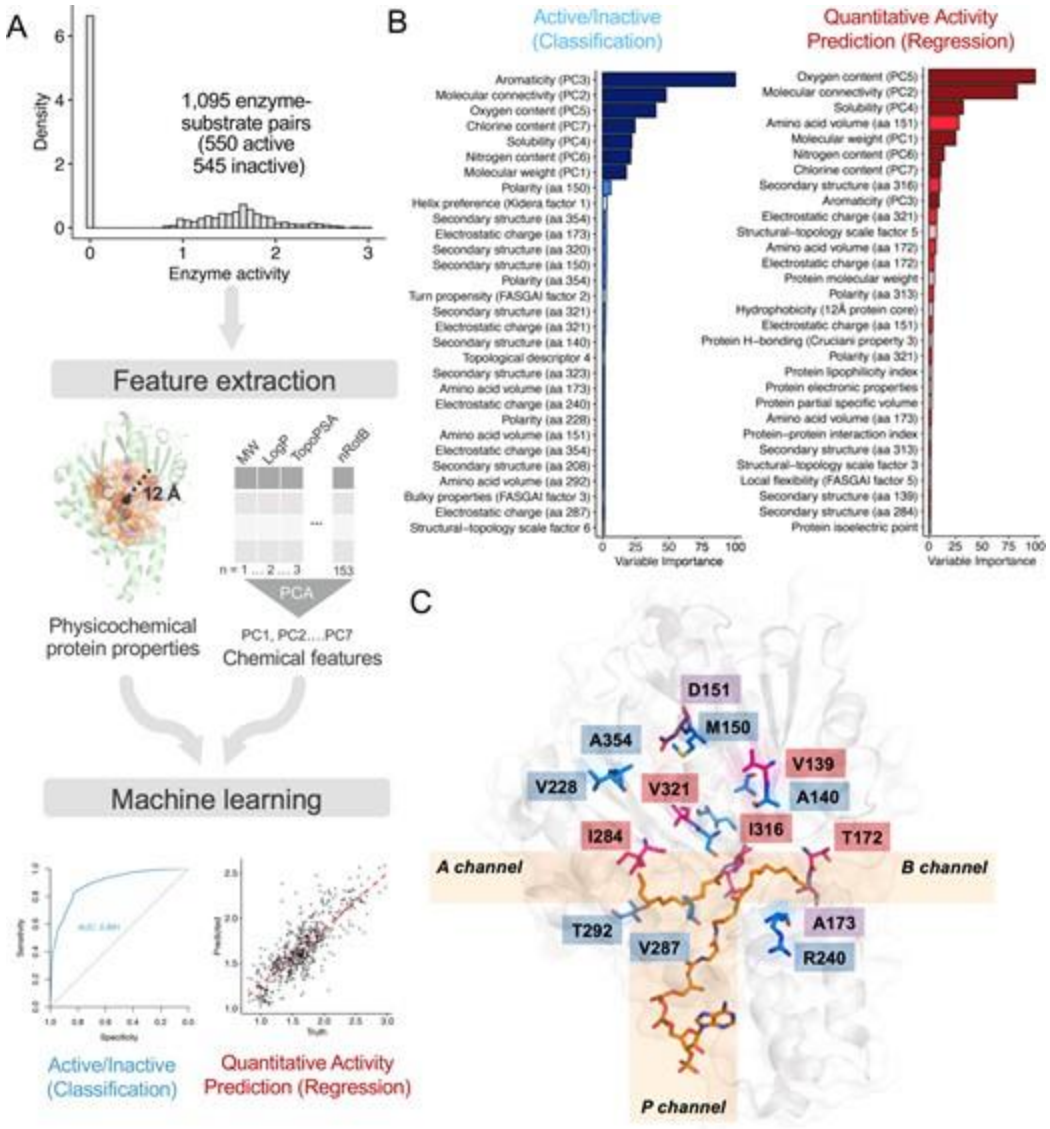


Figure 3-4. (A) Machine learning workflow used in this study. AUC, area under the receiver operating characteristic curve. (B) Variable importance scores for classification and regression models. (C) Important residues mapped onto the *X. campestris* structure (PDB ID: 4KU5) with fatty acid and acyl-CoA substrates bound. Residue colors correspond to variable importance in the classification model (blue), regression model (red) or both models (purple).

Machine learning algorithm	Training classification accuracy	Testing classification accuracy	Testing 95% confidence interval
Random forest	0.826	0.839	0.789–0.880
Feedforward neural network	0.732	0.777	0.722–0.826
Naïve Bayes	0.586	0.645	0.585–0.701

See Supplemental Figure 3-S3A for extended results.

Table 3-1. Machine learning classification results.

Notably, our machine learning algorithm identified channel residue 292 to be an important sequence feature for classification accuracy (Figure 3-4C). This residue had also been identified previously by BGA, further supporting a Thr in position 292 is likely important for substrate specificity. Moreover, a number of hydrophobic residues lining both binding pockets were identified by our model to be important including V287 and A173. We postulate these may assist in maintaining a hydrophobic environment within the binding pockets as was suggested in earlier crystallographic studies (31, 32). In *X. campestris*, an analysis of the residues lining substrate channels revealed the most common residues are Val, Leu and Ile (31). These aliphatic side chains promote binding of long-chain fatty acid substrates and likely play a role in determining which *p*NP substrates will bind and react in the OleA binding pocket.

Among the full-length protein indices included in our model, two protein indices for helix/turn propensity, Kidera and FASGAI (90, 91), were also important for prediction accuracy. We observed a negative correlation between the Kidera factor for helix/turn propensity of an enzyme and its average activity across all substrates (Pearson correlation = −0.23, *P*-value = 0.05). Overall, enzymes with lower helix propensity tended to be more active across all substrates, suggesting secondary structure flexibility enhances broad substrate

specificity. This is consistent with the importance of Val and Ile residues in our models since the flexibility of their aliphatic side chains likely allows a wider variety of substrates to bind.

'Pinch point' residues and substrate oxygen content influence enzyme–substrate regression models. We next examined whether we could quantitatively predict enzyme activity using regression models trained on the 550 active enzyme–substrate pairs. We evaluated three different machine learning algorithms (random forest, elastic net and multivariate adaptive regression splines). Again, random forest outperformed other models with a testing set R^2 of 0.75 (Table 3-2) and a testing RMSE of 0.243 ± 0.016 estimated from 1000 random training–testing dataset splits (Supplemental Figure 3-S3B).

Machine learning algorithm	Training RMSE	Training R^2	Testing RMSE	Testing R^2
Random forest	0.254	0.625	0.219	0.745
Multivariate adaptive regression splines	0.276	0.557	0.252	0.642
Elastic net	0.275	0.549	0.278	0.564

See Supplemental Figure S3B for extended results.
RMSE = Root mean square error.

Table 3-2. Machine learning regression results

Our regression models showed that PC5 (oxygen content) was the most important chemical feature for quantitative prediction of enzyme activity (Figure 3-4B). This is consistent with our findings that the *p*-nitrophenyl 2-(2-butoxyethoxy)acetate substrate had the highest average activity across all substrates and also has the highest number of oxygen atoms. PCs corresponding to molecular connectivity and solubility were the second and third

most important features, respectively. The molecular connectivity index (PC2) was positively correlated with average enzyme activity (Pearson correlation = 0.46, *P*-value < $2.2e^{-16}$). An examination of the loadings of PC2 further revealed positive associations between the number of atoms in the largest chain, the number of rotatable bonds and enzyme activity. These results further support the hypothesis for OleA-type thiolases preferring 'spacer' methylene carbons between the α-carbon and other functional groups such as phenyl, cyclic aliphatic, alkynyl or azido groups.

Residues 172, 173, 284, 287 and 316 were also identified as important by our regression algorithm and are known residues lining the substrate binding channels in the *X. campestris* structure (Figure 3-4C). Val 287 in the *X. campestris* structure plays a role in directing the path of the alkyl chain of the native substrate to curve around the hydrophobic Ile 345 side chain (31). Ile 284 is also in a critical position and likely interacts with Thr 292 based on the *X. campestris* crystal structure. A triad of residues (I284, A261 and T292) in channel A are predicted to form a 'pinch point' to mediate chain length specificity (32). Our results support this hypothesis and suggest these 'pinch point' residues are prime targets for rational design studies to alter thiolase substrate specificity.

Leave-one-out validation. To assess biases in our training set and determine how our models would perform with *p*NP substrates not included in model training, we performed leave-one-compound-out validation. We trained 15 separate models by omitting one compound completely from each model during training and then evaluating prediction accuracy for the held-out compound. Results from leave-one-compound-out validation for both classification and regression algorithms revealed that our model performed best with long-chain alkyl substrates including *p*-nitrophenyl dodecanoate and *p*-nitrophenyl 7-phenylheptanoate (Supplemental Figure 3-S5). Accuracy decreased for more chemically distinct and polar substrates such as *p*-nitrophenyl azidohexanoate and *p*-nitrophenyl 2-(2-butoxyethoxy)acetate. Overall, the leave-one-compound-out analysis indicated our models performed relatively well (>74% classification

accuracy) for most alkyl *p*NP substrates and underperformed with *p*NP compounds containing chemical moieties not included in the training set such as azides, ethers and trimethyl groups.

To test the phylogenetic bias of our model, we performed leave-one-taxon-out validation. OleAs in the gene library had been sampled from organisms belonging to 10 different taxonomic classes (Figure 3-2). We trained 10 separate models by withholding all sequences from organisms belonging to one taxonomic class and then testing prediction performance on the omitted class (Supplemental Figure 3-S5). We found that both classification and regression models performed more robustly with higher accuracy and RMSE values with leave-one-taxon-out validation than the leave-one-compound-out performance. This reflects that there is much more variability in average activity across different *p*NP compounds than across thiolases from different taxonomic classes. The two taxonomic classes which most significantly decreased classification and regression model performance when omitted were Gammaproteobacteria and Actinobacteria. This is consistent with the largest proportion of OleAs in the training set belonging to organisms from these classes.

Predictive web application. To make our machine learning models accessible to users with all levels of computational expertise, we created an interactive web application (z.umn.edu/thiolases). This web interface allows users to upload protein or nucleotide FASTA files of OleA-type thiolase sequences and uses machine learning models trained on the entire dataset to make rapid predictions for *p*NP substrate specificity. It provides predictions for activity with each of the 15 *p*NP substrates tested in this study as well as probability scores ranging from 0.5 (low confidence) to 1 (high confidence). Chemical structures of predicted substrates are displayed through the web interface (Supplemental Figure 3-S7) and results are downloadable in tabular format for further analysis.

Discussion

Over half of the thiolases tested were active with at least half of the *p*NPs tested. The remarkably broad substrate specificity in a large fraction of OleA-type enzymes in our library corresponds well with a previous study where hydrocarbons were extracted from twelve bacterial strains containing the complete *oleABCD* cluster (36). The authors reported that some bacteria, including strains from the genera *Xanthomonas* (Clade I) and *Kocuria* (Clade III), had OleABCD pathways that produced up to 15 different hydrocarbons identifiable by GC-MS (36). In our study, enzymes from *Xanthomonas* and *Kocuria* genera reacted with 11–14 different *p*NP substrates out of 15 tested. It is likely we have only sampled the 'tip of the iceberg' in terms of the number of potential substrates accepted by some highly active enzymes in this study such as the OleA from *K. sedentarius*.

In contrast, some enzymes tested such as the OleAs from *Shewanella putrefaciens* and *Psychromonas aquimarina* had narrow substrate specificity and weak activity. One explanation for the low activity of OleA enzymes from *Shewanella* and *Psychromonas* are their native preferences for a specific class of polyunsaturated fatty acid (Pfa) precursors that are produced by polyketide/fatty acid synthases encoded in a five-gene operon (*pfaABCDE*) upstream of the *oleABCD* cluster. Previously, Sukovich and colleagues extracted and verified a single 3,6,9,12,15,19,22,25,28-hentriacontanonaene product from the *Shewanella oleABCD* pathway (92). The authors postulated this product was derived from OleA-catalyzed condensation of two CoA-activated molecules of hexadeca-4,7,10,13-tetraenoic acid produced by the *pfa* operon. This was further supported by recent work in *Shewanella pealeana* demonstrating that OleA interacts with the Pfa synthase *in vivo* and mediates the transfer of Pfa precursors to the OleBCD complex to facilitate polyunsaturated hydrocarbon biosynthesis (93). We hypothesize here that the narrow substrate specificity observed in OleAs *Shewanella* and *Psychromonas* may be due to the co-evolution of *pfa* and *ole* sequences. To further investigate this, we identified 92 new OleA sequences from organisms that also encode *pfa* genes (Supplemental Material 3-S2). We used our machine learning models to make predictions for the

reactivity of these enzymes with 15 *p*NP substrates. We found that the predicted number of substrates accepted by OleAs co-localized with *pfa* genes was between 2 and 4 (average of 3), compared to an average of 8 substrates accepted by OleAs from organisms without *pfa* genes (Supplemental Table 3-S4, provided at 10.5281/zenodo.3743415). These predictions provide preliminary support for the hypothesis that *pfa*-associated OleA co-enzymes may have narrower substrate specificity than non-*pfa*-associated OleAs. Further research is required into how thiolases may have co-evolved with precursor biosynthetic enzymes such as Pfa synthases to affect OleA substrate selectivity.

The only substrate which was not active with any enzyme tested was *p*-nitrophenyl benzoate. The compound had been purchased, and we verified its purity by ^{1}H-NMR and mass spectrometry (Supplemental Figure 3-S1). We speculate the complete lack of activity with *p*-nitrophenyl benzoate may be due to the phenyl ring, lacking intervening methylene carbons, being bound in an unproductive manner such that the substrate carbonyl carbon is not accessible to the active site cysteine. Alternatively, steric hindrance at the α-carbon due to the bulky phenyl group could prevent cysteine attack. Other aromatic compounds tested, including *p*-nitrophenyl 3-(4-chlorophenoxy)propanoate, *p*-nitrophenyl 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoate and *p*-nitrophenyl 7-phenylheptanoate, were also relatively poor substrates compared to those with higher aliphalicity. In general, our data suggest that placement of aromatic groups further away from the α-carbon results in higher reactivity with OleA-family thiolases.

Based on our results, we infer the optimal *p*NP substrate for OleA enzymes has a side chain with at least 5–7 rotatable C–O or C–C bonds. Adequate spacing of 5–7 carbons between the α-carbon and other functional moieties such as alkynes, azides or cyclic aliphatic groups correlated with higher activity levels. Increasing the number of rotatable bonds was also positively associated with activity. Interestingly, the number of rotatable bonds in a molecule is considered to be a good descriptor of oral bioavailability of drugs (94). Overall, this work provides a foundation in the natural variation in OleA-

family thiolase substrate specificity to support future applications in synthetic biology such as using enzyme cascades to produce drug-like molecules. Recent advances have been made using biocatalytic cascades for the total synthesis of therapeutics such as the HIV drug, islatravir (95). We envision an expansion of the experimental and machine learning frameworks described here to aid in the design of enzyme and substrate libraries and ultimately use engineered thiolases for the production of bioactive molecules.

Substrates with terminal azide and alkyne moieties unexpectedly exhibited higher average reactivity with our enzyme library than the 'unmodified' alkyl *p*-nitrophenyl hexanoate and *p*-nitrophenyl heptanoate substrates. Substrates with azido- or alkynyl-groups were the second and third most reactive substrates screened, respectively. This represents the first experimental evidence of OleA-type thiolases reacting with substrates with 'clickable' functional groups. Copper(I)-catalyzed azide-alkyne cycloadditions, known as 'click' chemistry reactions, can be run under mild conditions that are particularly well-suited for biological systems. Recently, a reliable method for the synthesis of azides from primary amines opened new chemical possibilities for screening large libraries of clickable substrates (96). The ability for OleA-family thiolases to accept these compounds expands their applications in biological imaging and drug design through attachment of fluorophores or pharmacophores.

The overall goal of the assay was to provide a rapid screening method to identify thiolases with desired substrate specificity profiles that both express well and are active. One limitation of the whole-cell *p*NP assay is that the colorimetric readout from whole cells necessarily conflates protein expression with measured enzyme activity levels (43). All thiolases screened in this study were active with at least two *p*NPs (Supplemental Table 3-S2, provided at 10.5281/zenodo.3743415) suggesting all proteins were expressed. Since levels of protein expression can vary between experiments depending on exact induction time and cell density, we screened each of the active enzyme–substrate pairs in biological triplicates. However, we cannot rule out that the

measured weak activity of some enzymes is a result of consistently poor protein expression levels across multiple experiments.

Another limitation of the *p*NP assay is that it only measures the first step of the mechanism: transesterification. We are currently investigating Claisen condensation assay development since preferred substrates for the second step of the thiolase reaction (condensation) may be different than for transesterification. Previously, we purified several thiolases active with *p*NPs and measured their condensation activity with a variety of acyl-CoA substrates by GC-MS (43). However, this approach is not feasible for screening large libraries of enzymes. The development of a rapid assay to directly measure condensation activity will expand the applications of thiolases in synthetic biology.

In summary, we have quantitatively mapped the substrate specificity of thiolase enzymes through whole-cell assays, structural homology modeling, binding pocket analysis and machine learning. We surveyed a library of thiolase variants from taxonomically diverse organisms to assess natural variation in substrate scope and made predictions for new enzymes. This dataset will serve as a baseline for protein engineering studies to benchmark the performance of engineered enzymes against natural variants. Future experimental efforts will focus on altering key residues identified through our analysis to engineer thiolases to have high activity and desirable substrate specificity profiles.

The application of machine learning to predict substrate selectivity is generalizable to other enzyme families beyond thiolases and may be further improved through the use of ensemble or deep learning methods. While the feedforward neural network algorithm tested here did not outperform random forest, this is likely a result of our modest dataset size (1095 data points). We anticipate with larger datasets (>10 000 data points) that deep learning will outperform shallow learning algorithms like random forest. Overall, this work is a stepping stone towards a new frontier in biology where the combination of terabytes of meta'omic data with artificial intelligence can be used to learn complex patterns undetectable to the human eye. Here, we demonstrated a proof-of-principle for machine learning using physicochemical features to predict

enzyme substrate specificity. Ultimately, we envision experimental and computational methods such as those described here can be combined to advance the design of novel biological parts for the biosynthesis of high-value metabolites.

Acknowledgements

We acknowledge Yasuo Yoshikuni, Jan-Fang Cheng and Miranda Harmon-Smith at the U.S. Department of Energy Joint Genome Institute for their support, insights and discussions during construct design. Barbara Terlouw is acknowledged for critical feedback on the manuscript and on machine learning strategies with support from Janani Durairaj. We are also grateful to Romas Kazlauskas and Claudia Schmidt-Dannert labs for generous use of their lab equipment.

Funding

We thank the U.S. Department of Energy Joint Genome Institute for synthetic DNA. The work conducted by the U.S. Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, is supported under [DE-AC02-05CH11231]; The National Science Foundation Graduate Research Fellowship [00039202 to S.L.R.]; National Institutes of Health Biotechnology training grant [5T32GM008347-27 to M.D.S.]. We also acknowledge support from the MnDRIVE initiative for Industry and the Environment.

Conflict of interest statement. None declared.

CHAPTER 4 – *p*-Nitrophenyl esters provide new insights and applications for the thiolase enzyme OleA

"The real voyage of discovery consists not in seeking new landscapes, but in having new eyes." – Marcel Proust

This chapter is reprinted with permission from Elsevier *Computational and Structural Biotechnology Journal.*

CSBJ, 2021, 19, 3087-3096
DOI: 10.1016/j.csbj.2021.05.031

p-Nitrophenyl esters provide new insights and applications for the thiolase enzyme OleA

Megan D. Smith, Lambros J. Tassoulas, Troy A. Biernath, Jack E. Richman, Kelly G. Aukema, Lawrence P. Wackett

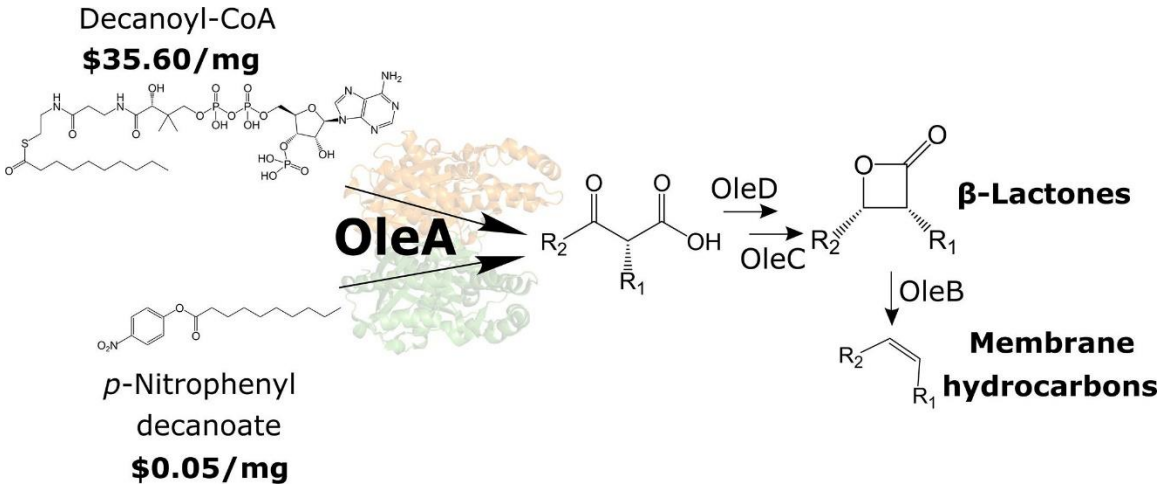
Summary

The OleA enzyme is distinct amongst thiolase enzymes in binding two long (≥C_8) acyl chains into structurally-opposed hydrophobic channels, denoted A and B, to carry out a non-decarboxylative Claisen condensation reaction and initiate the biosynthesis of membrane hydrocarbons and β-lactone natural products. OleA has now been identified in hundreds of diverse bacteria via bioinformatics and high-throughput screening using *p*-nitrophenyl alkanoate esters as surrogate substrates. In the present study, *p*-nitrophenyl esters were used to probe the reaction mechanism of OleA and shown to be incorporated into Claisen condensation products for the first time. *p*-Nitrophenyl alkanoate substrates alone were shown not to undergo Claisen condensation, but co-

incubation of *p*-nitrophenyl esters and CoA thioesters produced mixed Claisen products. Mixed product reactions were shown to initiate via acyl group transfer from a *p*-nitrophenyl carrier to the enzyme active site cysteine, C143. Acyl chains esterified to *p*-nitrophenol were synthesized and shown to undergo Claisen condensation with an acyl-CoA substrate, showing potential to greatly expand the range of possible Claisen products. Using *p*-nitrophenyl 1-^{13}C-decanoate, the Channel A bound thioester chain was shown to act as the Claisen nucleophile, representing the first direct evidence for the directionality of the Claisen reaction in any OleA enzyme. These results both provide new insights into OleA catalysis and open a path for making unnatural hydrocarbon and β-lactone natural products for biotechnological applications using cheap and easily synthesized *p*-nitrophenyl esters.

Graphical abstract

The thiolase enzyme OleA initiates the biosynthesis of β-lactone natural products and membrane hydrocarbons via acyl-Coenzyme A (CoA) substrates, but this study showed that *p*-nitrophenyl alkanoates can substitute, yielding new insights into the enzyme mechanism and providing a cheaper and more available starting substrate.



Introduction

Biohydrocarbons and β-lactone natural products molecules are of current interest in biotechnology. Hydrocarbons have been studied as petroleum product replacements and β-lactone natural products have demonstrated anti-obesity, anti-tumor, and antibiotic properties (4, 6, 44, 97–100). β-Lactones are analogs of β-lactams, both of which are chemical antagonists produced by living things, typically bacteria and fungi, to inhibit other organisms (101, 102). As a class, β-lactams are the major clinical antibiotics that have saved millions of lives but their effectiveness has been diminished in recent years by widespread pathogen resistance due to β-lactamases (103, 104). β-Lactones may fill in the some of the gaps for new antibiotics and anticancer treatments. For example, salinosporamide is in Phase three clinical trials for the treatment of multiple myeloma (21, 105). β-Lactones are end-products in some bacteria and metabolic intermediates in others. In the latter, they typically undergo decarboxylation to make olefinic hydrocarbons (27, 36). The intertwined pathways use homologous enzymes that are denoted as Ole, designating olefinic hydrocarbons, or they are named for their respective natural product; for example, LstAB in the case of lipstatin (12). OleA-type thiolase enzymes initiate biosynthesis of both the membrane hydrocarbons and many β-lactone natural products (4, 29).

OleA is a homodimer in the thiolase superfamily of enzymes and catalyzes a non-decarboxylative Claisen condensation of two acyl-CoA substrates consisting of C8-C16 carbon chains to produce hydrocarbons and natural products (29) (Figure 4-1A). Thiolase enzymes are involved in many different pathways in the metabolism of fatty acids, polyhydroxybutyrate storage, and natural product biosynthesis (47, 70, 106–108). OleA homologs that produce natural products include NltAB, a heterodimer biosynthesized by a *Nocardia* species that produces the β-lactone nocardiolactone (39) (Figure 4-1, *inset*). Another homologous protein pair, LstAB, is a heterodimer made by *Streptomyces toxytricini* that produces lipstatin, which is hydrogenated to make the antiobesity drug tetrahydrolipstatin, known commercially as Orlistat or Xenical (12, 40). The best studied enzyme catalyzing these condensation reactions with long chain

substrates is the OleA from the plant pathogen *Xanthomonas campestris* (29, 31–34).

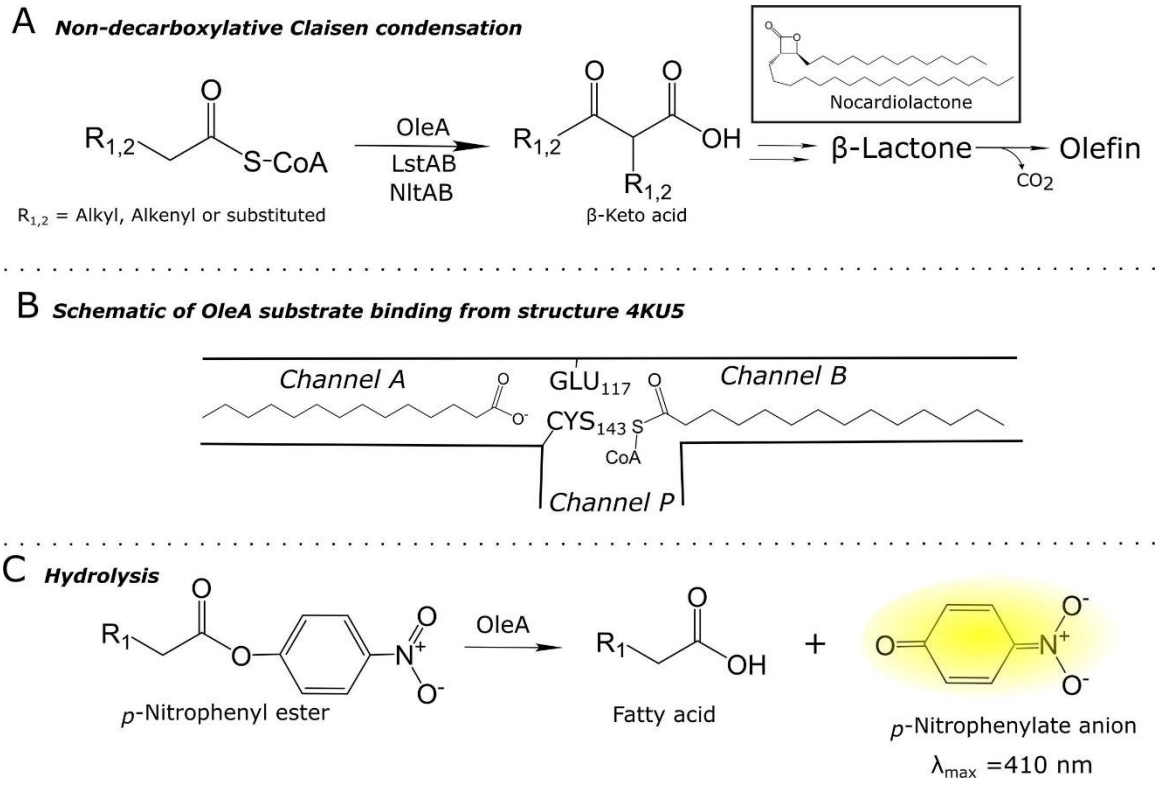


Figure 4-1. Reactions catalyzed by OleA and homologous enzymes. (A) Physiological reactions catalyzed by OleA and homologous enzymes LstAB and NltAB. (B) Schematic of OleA reaction showing three channel architecture of enzyme. (C) OleA screening reaction with *p*-nitrophenyl hexanoate making *p*-nitrophenol.

X-ray structures of the *X. campestris* OleA enzyme have been solved to 1.8 Å resolution, helping to reveal features of the Claisen condensation reaction (31–34). Initial studies with native enzyme and inhibitors revealed the reactivity of Cys143 and the occupation of a hydrophobic tunnel denoted as Channel A (Figure 4-1B). Structures were solved with irreversible inhibitors iodoacetamide and cerulenin, shown to be be covalently bonded to Cys143 with the hydrocarbon moiety extended toward and into Channel A, respectively.

Additional structures were determined with an inactive Cys143Ser mutant in which a fatty acid was bound in Channel A and an acyl chain covalently bonded to Coenzyme A were bound in Channel B and Channel P, respectively (Figure 4-1B). This was not a reaction complex since the fatty acid is not an activated intermediate but it showed the general binding mode of the two acyl chains and the Coenzyme A carrier in their respective channels. Further mutagenesis studies revealed that the Claisen reaction required E117 or, with lower activity, D117 (34). The carboxylate side chain is proposed to act as a general base to deprotonate a carbon adjacent to the acyl carbonyl carbon that acts as the Claisen reaction nucleophile.

Progress on OleA had initially been slowed by a cumbersome and time-demanding assay and was recently accelerated by the development of a high throughput readout of OleA activity, both *in vitro* and *in vivo* (43) (Figure 4-1C). The new assay was based on the observation that OleA catalyzes hydrolysis of *p*-nitrophenyl hexanoate at rates of hundreds of nmol per min, generating intensely yellow *p*-nitrophenol, that can be observed in seconds and determined quantitatively in a microtiter well-plate format. This assay has paved the way to conduct a broader palette of in vitro experiments with OleA. The method can also be used *in vivo* since *p*-nitrophenyl hexanoate diffuses into recombinant *E. coli* cells expressing OleA enzymes, while other cellular enzymes have relatively low activity with this substrate (43).

The new assay was leveraged, along with bioinformatics identifying thousands of putative OleA orthologs, to broaden OleA enzyme studies (43, 109). The method allowed identification of 72 additional OleA enzymes based on both bioinformatics criteria and *p*-nitrophenyl hexanoate hydrolysis activity (Figure 4-1C). Hexanoyl chains are shorter than those found in substrates preferred for Claisen condensation by the well-studied OleA from *X. campestris*. So it has remained an open question as to whether *p*-nitrophenyl esters could act as Claisen condensation donors or acceptors, or both.

The production of Claisen products starting with *p*-nitrophenyl esters rather than CoA esters would have numerous advantages in cost, ease of

substrate synthesis, and expanding the ability to generate chemical libraries of hydrocarbons and natural products. *p*-Nitrophenyl esters are typically <1% of the cost of corresponding CoA thioesters if purchased and they can be generated from thousands of available carboxylic acids via a high-yield synthetic procedure using *p*-nitrophenyl chloroformate (80, 109) (Supplemental Table 4-S1). With the availability of many cheap and easy-to-synthesize substrates for OleA, we could potentially make large product libraries to expand upon the range of products made by OleA enzymes. In a previous study, we found OleA enzymes that would react with more than one dozen *p*-nitrophenyl ester substrates containing aromatic, heterocyclic, and other functionalities that could provide useful pharmacophores if OleA would accept *p*-nitrophenyl-donated acyl groups and make Claisen condensation products.

The present study was conducted to explore the reactivity of *p*-nitrophenyl esters with OleA beyond the hydrolysis of *p*-nitrophenyl hexanoate (Figure 4-1C) to further probe the reaction mechanism and potentially produce new natural products (Figure 4-1A). The OleA from *X. campestris* was chosen for this study because the enzyme is significantly promiscuous in utilizing different substrates and X-ray structures are available. The results with this enzyme indicate that the *p*-nitrophenyl group can deliver an acyl chain into Channel A. A second *p*-nitrophenyl alkanoate is not competent for Claisen condensation but an acyl-CoA molecule can react to generate a mixed Claisen product. The ability to discriminate chains from discrete donors with a fixed channel occupancy allowed us to determine the directionality of the Claisen reaction for the first time. Previously undescribed chemical compounds were produced by co-reacting *p*-nitrophenyl and CoA esters. This study provided key insights into the OleA mechanism and opens the door for introducing unique functionality into the OleA condensation reaction to make new products of biotechnological importance.

Methods

Chemicals and Reagents. *p*-Nitrophenyl octanoate, decanoate, laurate, myristate, and palmitate, octanoyl-CoA, decanoyl-CoA, lauroyl-CoA, myristoyl-

CoA, palmitoyl-CoA, 10-nonadecanone, 12-tricosanone, and 14-heptacosanone were all obtained from Sigma-Aldrich (St Louis, MO). *p*-Nitrophenyl hexanoate was obtained from Tokyo Chemical Industry. *p*-Nitrophenyl heptanoate, nonanoate, 6-azidohexanoate, 3-(4-chlorophenoxy)propanoate and 1-^{13}C-decanoate were synthesized by the method of Engström *et al* [31]. Briefly, in a 10 mL flask, the respective carboxylic acid (0.5 mmol) and 4-nitrophenol chloroformate (0.5 mmol) were stirred under nitrogen 10 min. in cold (0 °C, 1 mL) anhydrous dichloromethane (DCM). A solution of 4-dimethylaminopyridine (0.05 mmol DMAP decarboxylation catalyst) and triethylamine (0.55 mmol) in 1.5 mL dry DCM was slowly added and stirred for 2 h. The reaction flask was nearly filled then with DCM and stirred rapidly with a drop of 6 N aq. HCl. The pH was adjusted to 6 (1 N NaOH), then the DCM phase and DCM washes were filtered through 2 g of silica gel collecting UV active eluate. The *p*-nitrophenol esters, analyzed by ^{1}H NMR and including *p*-nitrophenyl 1-^{13}C decanoate, were generally obtained in ≥98% purity and yield. The purity of *p*-nitrophenyl 1-^{13}C-decanoate used in this study was 90–95%.

Bacterial strains, vectors and genes encoding OleA wild-type and mutant proteins. All genes used in this study were for *Xanthomonas campestris* Ole A wild-type (NP_635607.1) and mutant derivatives C143S, C143A, and T292M. All OleA genes were expressed in *Escherichia coli* T7 Express cells from vector pET28b+ with a T7 promoter. The T292M mutant OleA gene was obtained from the Department of Energy's Joint Genome Institute (JGI) and expressed in the same vector. All contained an N-terminal 6x His-tag. *E. coli* cells were grown to an OD of 0.4 and induced at 16 °C overnight prior to protein purification.

Protein purification and handling. *Xanthomonas campestris* Ole A wild-type (NP_635607.1) and mutant derivatives C143S, C143A, and T292M were lysed by French Press and proteins purified by the general method described by Frias *et al* (29) using Ni-column chromatography. The column was washed with a buffer of 20 mM Tris·HCl at pH 7.5 containing 500 mM NaCl. The column was

subsequently washed with two column volumes each of 30 mM, 100 mM and 400 mM imidazole in the same buffer. OleA eluted half-way through the last 400 mM column wash. Protein was then eluted through a Sephadex G-10 column to remove imidazole that can catalyze hydrolysis of *p*-nitrophenyl esters. Protein was then flash frozen and stored at −80 °C until use. Yields of protein were typically 15–20 mg per liter. SDS-PAGE showed protein homogeneity as previously reported (29). Mutant proteins were purified via the same methods and showed similar behavior upon handling.

General assay for *p*-nitrophenyl ester hydrolysis. 50 mM Tris·HCl pH 8.0 was added to a 96-well Flat Bottom Suspension Culture Plate (Cat#:25–104 Genesee Scientific) containing 30 µg of OleA C143S or C143A, 5% ethanol, and 200 µM *p*-nitrophenyl hexanoate in a 200 µL total volume. A SpectraMax Plus 384 Microplate Reader (Molecular Devices, San Jose, CA) was used, and *p*-nitrophenol absorbance was read at 410 nm. Absorbance was read at 410 nm every 5 min for at least 2 h at 37 °C. A standard curve was developed in a buffer containing 50 mM Tris·HCl at pH 8.0. The extinction coefficient for *p*-nitrophenol was determined to be 15,548 $M^{-1}cm^{-1}$, and the path length through the 200 µL liquids in microtiter wells was experimentally determined to be 0.58 cm. All reactions were run in parallel on the same microtiter plates with 3–5 replicates for each data point. Controls of each *p*-nitrophenyl alkanoate chain length containing no protein were used to correct for any non-enzymatic hydrolysis in buffer. For wells containing wild type OleA, only 4 µg of protein was added to the wells, and measurements were taken every minute for 30 min. Other mutant and inhibited assays were varied as described below.

***p*-Nitrophenyl ester hydrolysis with inhibited or mutant enzymes.** The assays were run as generally described for wild-type OleA with the following changes. For assays containing inhibited wild type enzyme or C143A/S mutants, 30 µg of OleA was used, and absorbance was determined over a 2 hr time course. For assays with wild-type enzyme inhibited with cerulenin and

iodoacetamide, 1.25 mM of the inhibitor was added to 30 µg OleA enzyme and buffer and allowed to react at room temperature for 1 hr prior to addition of *p*-nitrophenyl hexanoate.

For reactions comparing OleA wild type and T292M mutants, reactions were run in parallel. Both sets of reactions tested C_6, C_7, C_8, C_9, C_{10}, C_{12}, and C_{14} alkanoate esters of *p*-nitrophenol. The buffer, pH and other parameters were as described for the general assays. Reactions contained 6 µg of enzyme and were initiated by the addition of the respective *p*-nitrophenyl ester.

GC–MS conditions for screening Claisen condensation activity with *p*-nitrophenyl esters. Gas chromatography-mass spectrometry (GC–MS) was carried out on an Agilent 7890a gas chromatograph and an Agilent 5975c mass spectrometer (Santa Clara, CA). 500 µL reactions were run using 8 µg OleA, 5% ethanol, and 200 µM *p*-nitrophenyl esters in Tris·HCl 50 mM pH 8.0. The reaction was allowed to proceed overnight at room temperature. The product was extracted into 500 µL of methoxy-*t*-butyl ether (MTBE). Methylation was performed using 100 µL diazomethane in diethyl ether. The resultant product was injected into the gas chromatograph using both flame ionization (FID), and electron impact mass spectrometry detection at 70 V. The inlet temperature was 250 °C. The elution program was as follows: hold 60 °C for 8 min followed by a 15 °C increase per minute until 320 °C was attained and held constant for 5 min.

The products of the Claisen condensation are β-keto acids that completely decarboxylate in the gas chromatograph to yield ketones (29). The resultant ketones of different molecular weights are thermostable under the GC regime described above and yield very characteristic mass spectra. To determine the lowest level of activity that was detectable in the assay, standard curves were obtained for commercially-obtained 10-nonadecanone, 12-tricosanone, and 14-heptacosanone via GC–MS (Supplemental Figure 4-S1-3). Standards were dissolved in MTBE at 400 nM, 250 nM, 200 nM, 150 nM, 100 nM, 50 nM concentrations and 1 µL standards were run using the same protocol detailed in the GC–MS condensation assay. Ketone peaks were recorded at 17.1 min for

10-nonadecanone, 19.6 min for 12-tricosanone, and 21.9 min for 14-heptacosanone. Peaks were integrated to generate the standard curves.

***p*-Nitrophenyl 1-^{13}C-decanoate and myristoyl-CoA used to investigate the Claisen nucleophile.** 500 µL reactions were run using 8 µg OleA, 5% ethanol, 100 µM *p*-nitrophenyl 1-^{13}C-decanoate and 100 µM myristoyl-CoA in Tris·HCl 50 mM pH 8.0. The reaction was allowed to proceed for 30 min. To detect the intact β-keto acid products, the enzyme reactions were extracted with MTBE after 30 min, the product(s) methylated by diazomethane in ether via the standard extraction protocol, and subjected to GC–MS as previously described. An unlabeled *p*-nitrophenyl decanoate substrate control was run in parallel at the same time. The mass spectra are presented in the results section. In separate experiments, the β-keto acid decarboxylation products, aliphatic ketones, were analyzed omitting the methylation step and carrying out GC–MS directly. The half-life of decarboxylation of the β-keto acids to ketones was previously determined to be ~8 h and, if not methylated, they undergo complete decarboxylation in the GC injection inlet (29).

Computational methods. The OleA T292M variant was modelled using the Mutagenesis tool in PyMOL, Version 2.0 (Schrödinger) as shown in Figure 4-2. The specific rotamer chosen for the methionine variant came from a backbone-dependent rotamer library and was the only rotamer that could directly add steric effects to substrate binding. Modelling of the enzyme thioester intermediate was done by placing the myristyl carbonyl carbon linked to the sulfur atom of OleA residue C143 in the existing ligand electron density from structure PDB 4KU3 using Coot, v0.8.9.2 (110). The carbonyl moiety of the thioester intermediate was positioned manually into the oxyanion hole formed by H285 and N315 and the OleA-thioester complex was refined using default restraints using the Refmac5 tool within the software, CCP4, version 7.0 (111). The refinement minimized the energy of the bonded and non-bonded interactions of the thioester intermediate

which resulted in the final model shown in Fig. 6. Chemical substances were searched using SciFinder (112, 113).

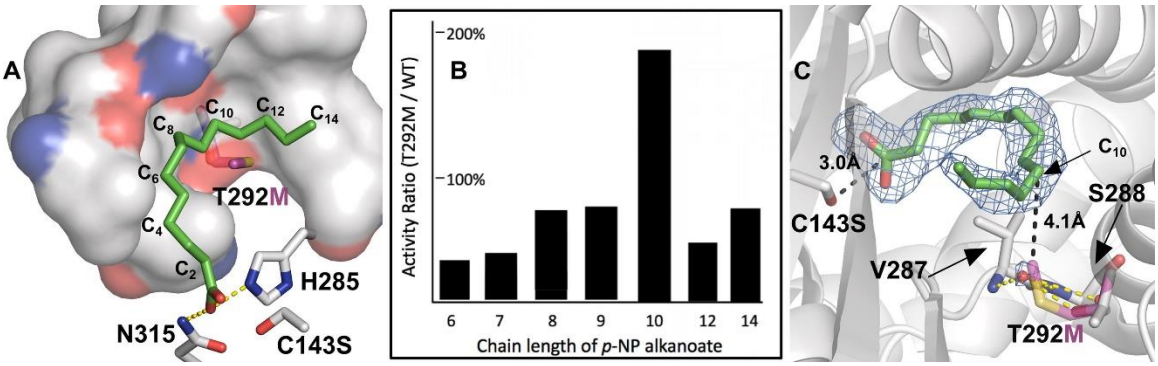


Figure 4-2. Mutation in Channel A effects length of p-nitrophenyl ester chain accepted. (A) The “bend” of myristic acid bound in OleA channel A. Surface representation of the channel A cavity with the T292M variant modelled. The carboxylic acid moiety of myristic acid is proximate to the catalytic cysteine, C143, and H285, N315 that form the oxyanion hole. The C143S variant inactivates OleA and allows for co-crystallization with myristic acid (PDB 4KU3) (B) Activity of the T292M variant over wild-type OleA versus *p*-nitrophenyl ester chain length. Activity measured as described in the Methods section (C) Modeled binding mode of T292M variant. Channel A of OleA with a Fo-Fc map contoured at 3σ carved around myristic acid bound in the channel (PDB 4KU3). The T292M variant when modelled may occlude a highly coordinated water molecule (red sphere) and have increased Van der Waals interactions with a C_{10} *p*-nitrophenyl alkanoate. In the wild-type, the hydroxyl of T292 and the backbone atoms of V287 and S288 coordinate the water molecule and occlusion of binding this water is proposed to alter the channel A cavity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

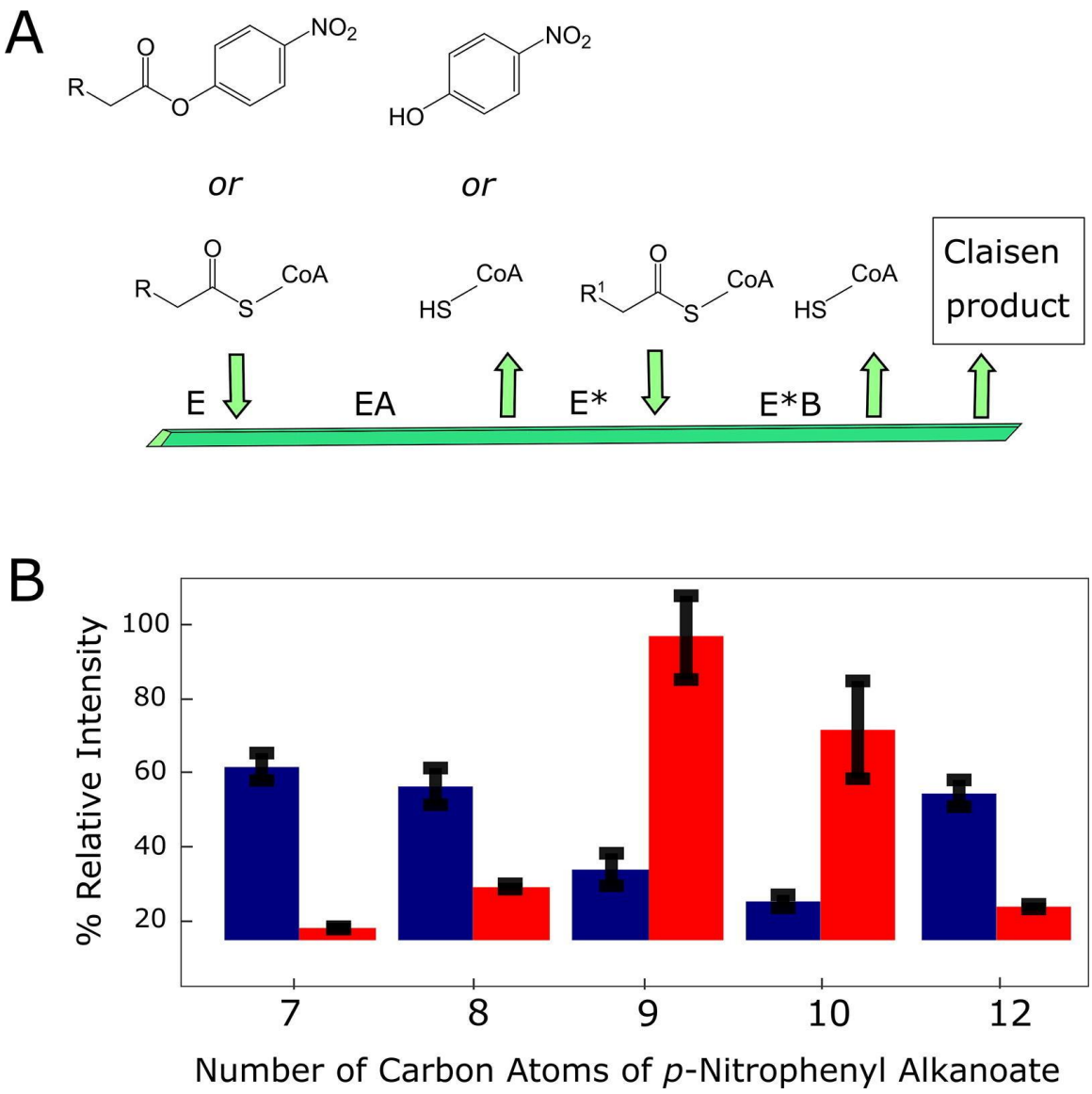


Figure 4-3. OleA catalyzes reaction between p-nitrophenyl ester and CoA ester. (A) Cleland diagram illustrating ping-pong mechanism in which transesterification from a *p*-nitrophenyl or CoA ester leads to an identical intermediate E* enzyme state. (B) Bar graph showing relative products from condensation of two CoA esters (blue) and a CoA ester and a *p*-nitrophenyl ester (red). The CoA ester was myristoyl-CoA and the *p*-nitrophenyl ester chain length was varied as shown on the X-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Results and discussion

No Claisen condensation with *p*-nitrophenyl esters alone. *p*-Nitrophenyl esters react rapidly with OleA to release *p*-nitrophenol and this has been used to identify OleA enzymes in dozens of different bacteria and to probe OleA substrate specificity (43, 109). Here, we wished to produce Claisen condensation products using *p*-nitrophenyl esters. In the present study, *p*-nitrophenyl esters were reacted in sixteen different combinations and analyzed for Claisen product formation. All combinations of straight-chain C_9, C_{10}, C_{12} and C_{14} alkanoic acid esters were examined. No Claisen products were detected in any reaction. Available standards for the analyzable Claisen condensed products were run in parallel. The standards were readily detected at a level of 35 pmol (Supplemental Figure 4-S1 – 4-S3). Positive controls with C_{10}, C_{12} and C_{14}-CoA thioesters, the physiologically relevant substrates (36), gave expected levels of Claisen condensation products.

***p*-Nitrophenyl ester reactions require the OleA active site cysteine.** With wild type *X. campestris* OleA, acyl-CoA substrates undergo acyl group transfer to Cys143 and subsequently two reactions: (i) a Claisen condensation with an acyl chain bound in another channel or (ii) a hydrolysis of the acyl enzyme to generate a fatty acid (29). The ratio of the reactions varies with the length of the acyl chain and reactions occur on a time scale of minutes. In some cases, the hydrolysis product predominates over the Claisen product. OleA single-site mutants, C143S or C143A, do not undergo Claisen condensation (31). Additionally, the C143 mutants catalyze a very slow hydrolysis of acyl-CoA thioesters, requiring overnight incubations to observe significant fatty acid product (31). While rates were too slow to measure precisely, they are estimated to be less than one percent of the rates of the wild-type Claisen and hydrolysis reactions, consistent with the C143 acylation reaction preceding both Claisen condensation and hydrolysis of CoA thioesters (31). We believe this is likely due to direct hydrolysis by a water molecule. Also, consistent with the role of C143, iodoacetamide and cerulenin were shown to be potent irreversible inhibitors of

OleA and X-ray studies revealed they inhibit by covalently modifying the cysteine (32).

In the present study, the significant reaction of *p*-nitrophenyl esters with OleA was seen to similarly be dependent on C143. Reaction of the wild-type enzyme with iodoacetamide and cerulenin virtually eliminated the hydrolysis reaction, determined here at <0.1% of the rate of uninhibited enzyme (Table 4-1; Supplemental Figure 4-S4). Similarly, C143S and C143A mutants were severely impaired with the rates of hydrolysis <1% of the wild-type rate (Table 4-1; Supplemental Figure 4-S5). The alkyl chain of cerulenin, following cysteine alkylation, was previously shown by X-ray crystallographic studies to occupy what has been denoted as the Channel A, in which the first acyl substrate binds to OleA (32). These studies are consistent with a transfer of alkyl chains from *p*-nitrophenyl esters to C143 with the hydrophobic alkyl moiety extending into Channel A.

Enzyme form	**Standard assay or additive**	**Specific Activity (nmol/min per mg)**	**k_{cat} (min^{-1})**
Wild-type	Standard	270 ± 0.05	10
Wild-type	Iodoacetamide	0.16 ± 0.08	0.006
Wild-type	Cerulenin	0.26 ± 0.09	0.010
Mutant, C143A	Standard	0.75 ± 0.08	0.029
Mutant, C143S	Standard	0.69 ± 0.08	0.026

Table 4-1. Enzyme activity measured with *p*-nitrophenyl hexanoate as described in Methods for wild-type enzyme, wild-type enzyme reacted with Channel A inhibitors and mutant enzymes.

Mutation in Channel A effects length of *p*-nitrophenyl ester chain accepted. Similarities in the OleA-catalyzed hydrolysis reactions with acyl-CoAs and acyl-*p*-nitrophenyl esters suggested that both undergo transesterification to Cys143 with the acyl group occupying Channel A. To further investigate these commonalities, we made mutations in Thr292 that forms part of the Channel A binding site. Here Thr292 was mutated to a methionine, a change expected to alter the binding and reactivity of *p*-nitrophenyl alkyl esters (Figure 4-2A).

The T292M mutant was stable and could be purified in reasonable yield and showed clearly altered chain selectivity of for *p*-nitrophenyl alkyl esters, further supporting the conclusion that *p*-nitrophenyl alkanotes occupy Channel A. The wild type enzyme has the lowest activity with *p*-nitrophenyl decanoate in a C6 through C12 panel of substrates (29) whereas the T292M mutant showed comparatively high activity with *p*-nitrophenyl decanoate (Figure 4-2B). Indeed, while the mutant is lower in activity against all other chain lengths, it shows nearly double the activity of the wild-type with the C10 alkanoate ester. Mutant OleA enzymes with larger, more hydrophobic side chains, T292I and T292V and T292F, were tested in lysed cells and failed to show any significant hydrolysis activity with either *p*-nitrophenyl hexanoate or *p*-nitrophenyl dodecanoate. These larger hydrophobic side-chains may result in the Channel A not opening sufficiently to accommodate a *p*-nitrophenyl ester.

X-ray structures with C_{12} and C_{14} acyl chains bound show a distinct bend in the chain between carbon atoms C_8-C_{11}, the region of T292 (Figure 4-2A). Note that the chain is likely displaced further down the channel in the X-ray structure PDB 4KU3 that contains a bound carboxylic acid. In a true reaction complex, for which no experimental structure exists currently, the acyl chain would be directly bonded to C143, eliminating the second oxygen atom of the carboxylate. This would effectively position the chain such that the C_9-C_{10} atoms start the bend region depicted in Figure 4-2A and 4-2C. In that region of Channel A, T292 participates in coordinating a tightly bound water, shown as a red sphere, along with the backbone atoms of V287 and S288. Our modeling indicates that all these interactions would be disrupted by a methionine residue in

position 292. We propose that the T292M mutation displaces the bound water and the methionine can provide a favorable Van der Waals interaction with the terminal methyl group of the C_{10} *p*-nitrophenyl substrate. This configuration would be consistent with favorable binding of a C_{10}-chain with the mutant compared to a more unfavorable binding by the wild-type, and this is what was observed experimentally.

Mixed *p*-nitrophenyl ester and CoA-thioester reactions. The present study suggested that *p*-nitrophenyl acyl chains bind in Channel A prior to hydrolytic cleavage, and previous data support the OleA-catalyzed Claisen reaction following a ping-pong mechanism like other thiolases (32, 47, 70, 106, 107). Via that mechanism, acyl transfer from the *p*-nitrophenyl ester to C143 is followed by leaving of the first product, *p*-nitrophenol (Figure 4-3A). This intermediate state would be identical to the covalent enzyme intermediate (E* in Figure 4-3A) formed from acyl transfer from acyl-CoA and leaving of CoA. This led us to the hypothesis that a subsequent acyl-CoA substrate can occupy Channel B and undergo Claisen condensation with the Channel A acyl chain, regardless of whether the first chain comes from a CoA or a *p*-nitrophenol donor.

This was tested with mixed *p*-nitrophenyl alkanoate + acyl-CoA incubations in which the chain lengths differ, allowing differentiation from condensation reactions occurring from two acyl-CoA substrates of the same chain lengths. Our choice of *p*-nitrophenyl ester chain lengths was guided by two factors: (1) much higher solubility of shorter chain length *p*-nitrophenyl esters (Supplemental Table 4-2S), and (2) previous observations that Claisen reactions occur with intermediate chain length acyl-CoA substrates (C_8-C_{12}). In this context, we tested C_7, C_8, C_9, C_{10}, and C_{12} straight chain alkyl esters of *p*-nitrophenol, each separately in a mixture with a preferred chain-length second substrate, myristoyl-CoA (C_{14}).

The mixtures were allowed to react until completion, extracted with methyl-*t*-butyl ether, and analyzed by GC–MS for the Claisen product. There was extensive hydrolysis but significant Claisen products were observed, including

mixed Claisen products derived from one *p*-nitrophenyl ester and the CoA thioester (Figure 4-3B; Supplemental Figure 4-S6). When examining the mixed condensation products only, product derived from a *p*-nitrophenyl nonanoate had the greatest relative amount observed (100%), while the product derived from *p*-nitrophenyl decanoate also yielded a substantial amount (70%) (Figure 4-3B). Other condensed products were significantly less, due largely to hydrolytic cleavage of the *p*-nitrophenyl ester. These data are consistent with *p*-nitrophenyl acyl chains binding in Channel A, with a kinetic competition between hydrolysis of the enzyme C143 thioester intermediate and acyl-CoA binding in Channel B to allow for the Claisen condensation to proceed.

Mixed incubations of *p*-nitrophenol [^{13}C]-ester and CoA thioester. The mixed Claisen reaction could be occurring in either of two ways. In one way, the α-carbon of the acyl chain delivered by *p*-nitrophenol to C143 and in Channel A is activated to attack the carbonyl of the chain bound in Channel B (Figure 4-4A, left). In the second way, the α-carbon of the acyl chain covalently attached to coenzyme A and in Channel B is activated for the Claisen condensation (Figure A, right). To differentiate between those mechanisms, it was necessary to use different chain length *p*-nitrophenyl and CoA esters with one chain labeled. We could then analyze the unstable mixed Claisen product β-keto acid directly and, alternatively after decarboxylation, both by mass spectrometry. To that end, *p*-nitrophenyl 1-^{13}C-decanoate was synthesized and a protocol was developed to methylate the β-keto acid using diazomethane to stabilize it for GC–MS analysis. We then conducted an experiment with *p*-nitrophenyl 1-^{13}C-decanoate and myristoyl-CoA, analyzing the C_{24} beta-keto acid thus formed. A control without ^{13}C-labeling was conducted in parallel for direct comparison.

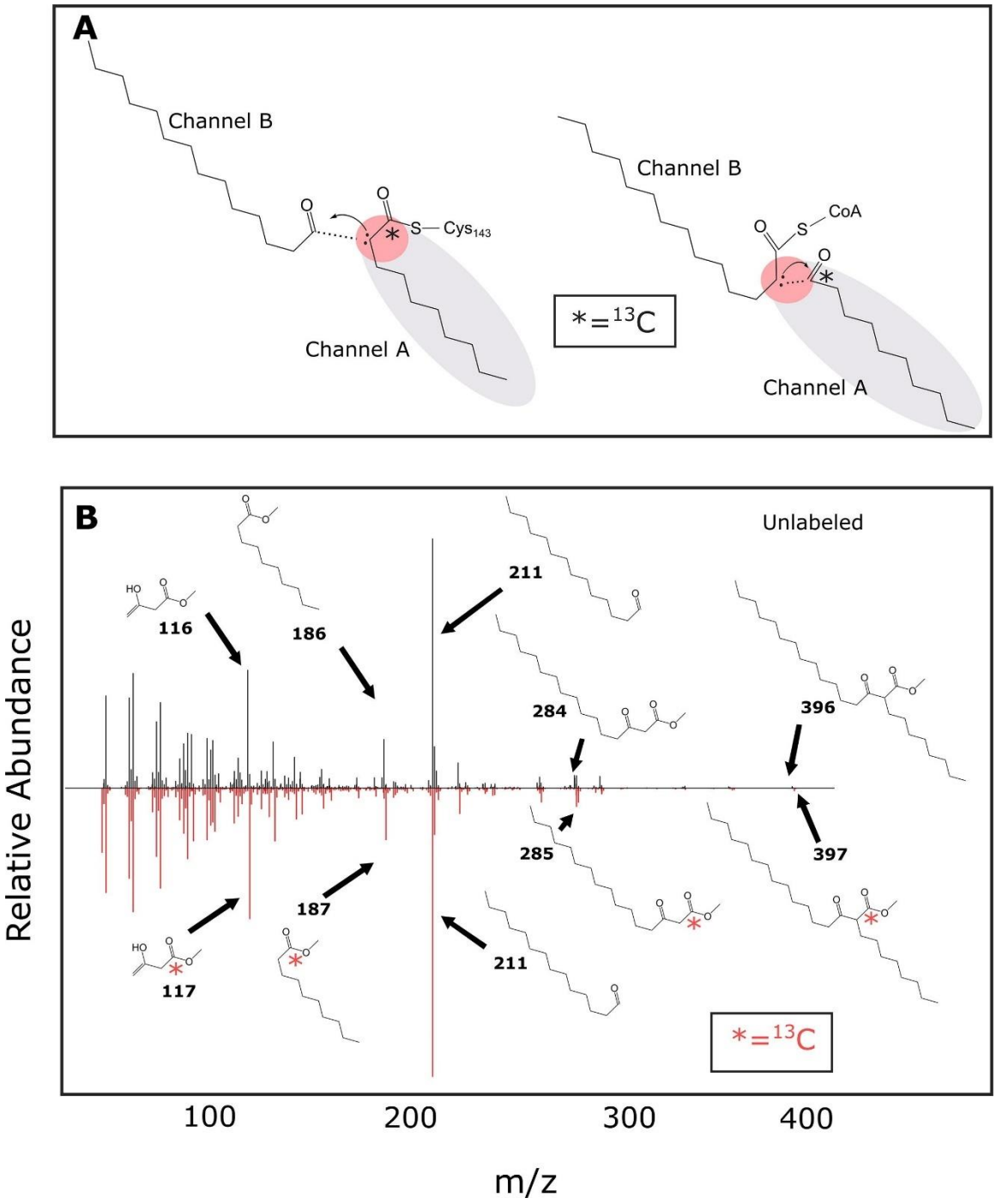


Figure 4-4. Directionality of OleA Claisen reaction demonstrated with a 1-^{13}C-p-nitrophenyl decanoate that only reacts from binding in Channel A. (A) The two possible mechanisms of chain activation and C–C bond formation in *X. campestris* OleA. (B) Mass spectrum of the methylated β-keto acid product emanating from the condensation of *p*-nitrophenyl 1-^{13}C-decanoate and myristoyl-CoA. (red), or p-nitrophenyl decanoate and myristoyl-CoA (black).

Major fragments of the β-keto acid product are depicted, along with their mass and arrows pointing to their corresponding peak. A red asterisk present indicates the ^{13}C label. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The mass spectrum of 2-octyl-3-ketohexadecanoate methyl ester showed a small parent ion that contained only one ^{13}C-label (Figure 4-4B). The mass, *m*/*z* = 397, was consistent with a β-keto acid of 24 carbons plus the methyl group from diazomethane. This indicated that the product was formed from one C_{10}-*p*-nitrophenyl ester and one C_{14}-CoA thioester. The fragmentation pattern further revealed that the ^{13}C -label was found exclusively in the carboxyl group of the product. Typically, ketones fragment at the bond adjacent to the carbonyl carbon, as is observed here. Moreover, unlabeled and ^{13}C-labeled mass spectra are highly comparable, with exception of a mass shift of one for the identified fragments. The major fragment with *m*/*z* = 211 is consistent with a fragment of $CH_3(CH_2)_{12}C(O)$ derived from myristoyl-CoA and that is not expected to shift when using ^{13}C-1-decyl-*p*-nitrophenol based on the mechanism shown in Fig. 4A, on the left. The corollary to this is that fragments containing the carbonyl carbon of *p*-nitrophenyl 1-^{13}C-decanoate would be expected to shift up by one mass unit and this is observed in Figure 4-4B.

In a separate experiment, we reacted *p*-nitrophenyl 1-^{13}C-decanoate and myristoyl-CoA and directly chromatographed the product via GC–MS. β-Keto acid products are known to rapidly decarboxylate in the heat of the GC (29). If the ^{13}C -label is in the carboxyl group, that will be lost as carbon dioxide and the mass spectra of the ketone products from ^{13}C-labelled and nonlabelled substrates would be identical. Indeed, we observed identical mass spectra, no hint of a mass shift of all identifiable fragments (Supplemental Figure 4-S7). Collectively, these data demonstrated that the chain from the *p*-nitrophenyl ester that is tethered to the cysteine and bound in Channel A is activated to carry out the Claisen reaction as shown in Figure 4-4A, left.

Generating novel OleA products. The OleA-catalyzed *p*-nitrophenyl ester hydrolysis reaction has, in the past, proven useful mainly for screening purposes, but it can now be extended with the new insights obtained in this work. Novel *p*-nitrophenyl acyl chains can be introduced into Claisen products because they uniquely go into Channel A and condense with a CoA ester bound in Channel B. This greatly expands the range of Claisen products that can be generated by OleA enzymes, which have been shown to have a promiscuous Channel A binding site.

Moreover, the mixed Claisen products can be generated more efficiently and inexpensively via *p*-nitrophenyl esters in place of CoA esters. Acyl-CoA thioesters are expensive to purchase, > $30,000 per gram (Supplmental Table 4-S1), and their synthesis requires a non-ideal coupling of the highly water soluble CoA moiety with an activated carboxylic acid that can hydrolyze in water, making for low to modest yields of product (114). Alternatively, fatty acid CoA ligase enzymes and ATP can be used to biosynthesize specific CoA esters (115, 116). By contrast, *p*-nitrophenyl esters can be made in high yields using *p*-nitrophenyl chloroformate in direct reaction with thousands of cheap and commercially available carboxylic acids, opening the door to making novel substrates for OleA-catalyzed reactions. Here we synthesized 4-nitrophenyl 3-cylopentylpropionate and 4-nitrophenyl 6-heptynoate using a facile one-pot reaction as previously described (80, 109). These compounds were predicted to bind in Channel A and undergo reaction with co-incubated myristoyl-CoA substrates that would bind in Channel B.

The compounds reacted with OleA to form Claisen products that are derived from one each of the *p*-nitrophenyl donors and myristoyl-CoA as the second substrate (Figure 4-5). GC–MS was used to separate and identify products via mass fragments. The β-keto acid products are unstable in the gas chromatograph and undergo decarboxylation to yield ketones, as illustrated in Figure 4-5. Unique peaks were observed in each OleA reaction mixture that were absent in chromatograms of incubations lacking OleA. Each product was a

distinct ketone with 20 or 21 carbon atoms, which was expected to elute in the 17–20 min range, as was observed here. Each compound was subjected to electron impact mass spectrometry. The cyclopropyl ketone and alkyne ketone mass spectra both showed highly diagnostic fragmentation patterns (Figure 4-5 insets and top). Neither of the ketone compounds, nor the β-keto carboxylic acids from which they are derived, have been described in the chemical literature. Their novelty was determined by searching SciFinder, which contains more than one hundred million chemical substances and is regularly updated (112, 113).

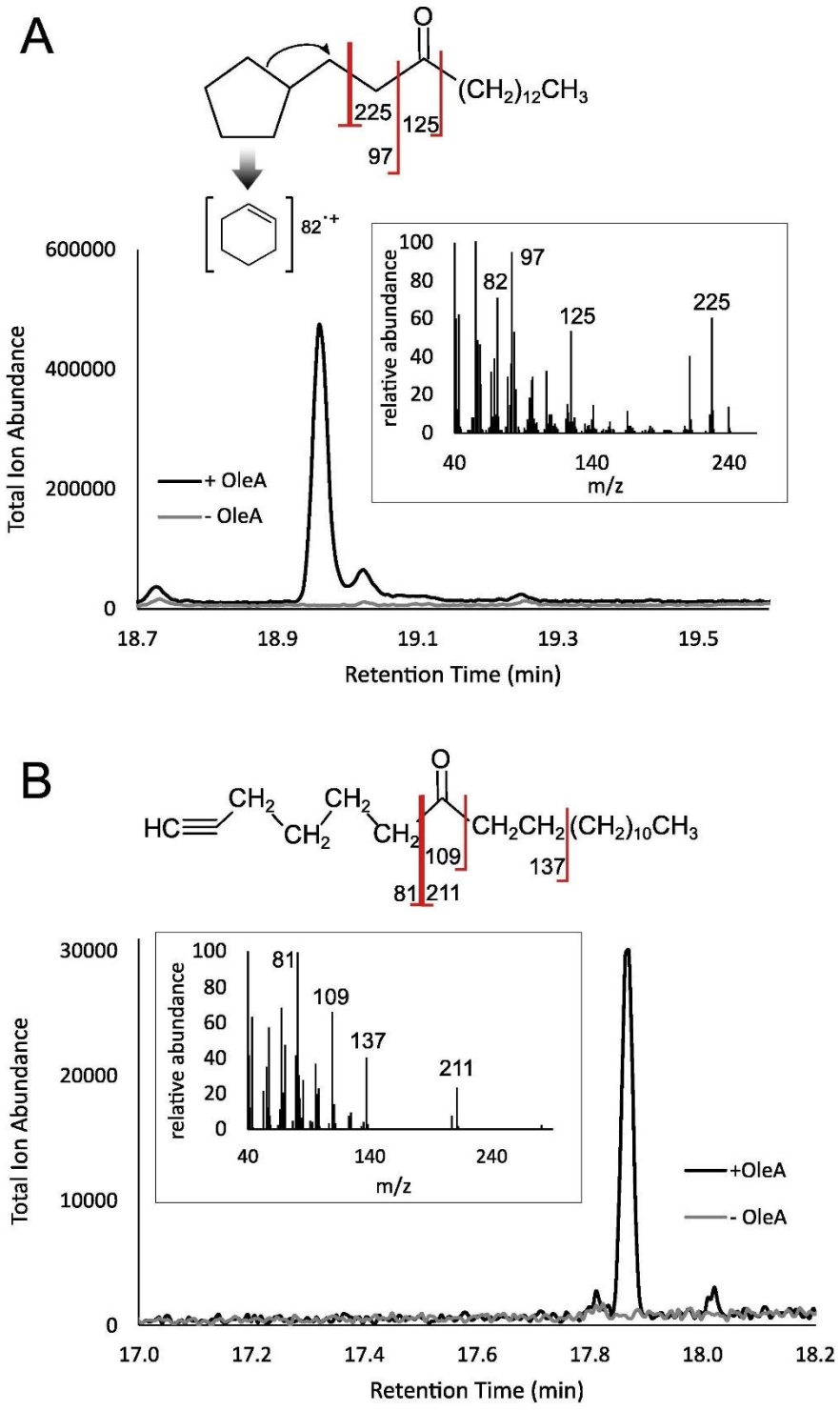


Figure 4-5. Identification of OleA-catalyzed Claisen condensation products combining acyl groups delivered by p-nitrophenol and CoA, respectively. The CoA ester was myristoyl-CoA condensed with either *p*-nitrophenyl 3-cyclopentylpropionate (A) or *p*-nitrophenyl 6-heptynoate (B). Both (A) and (B)

show gas chromatograph traces of the condensed product (+OleA) and a no enzyme control (- OleA). The inset boxes are electron impact mass spectra of the major peaks. Product structures are shown above that. Red bars and numbers indicate mass fragments (*m*/*z*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

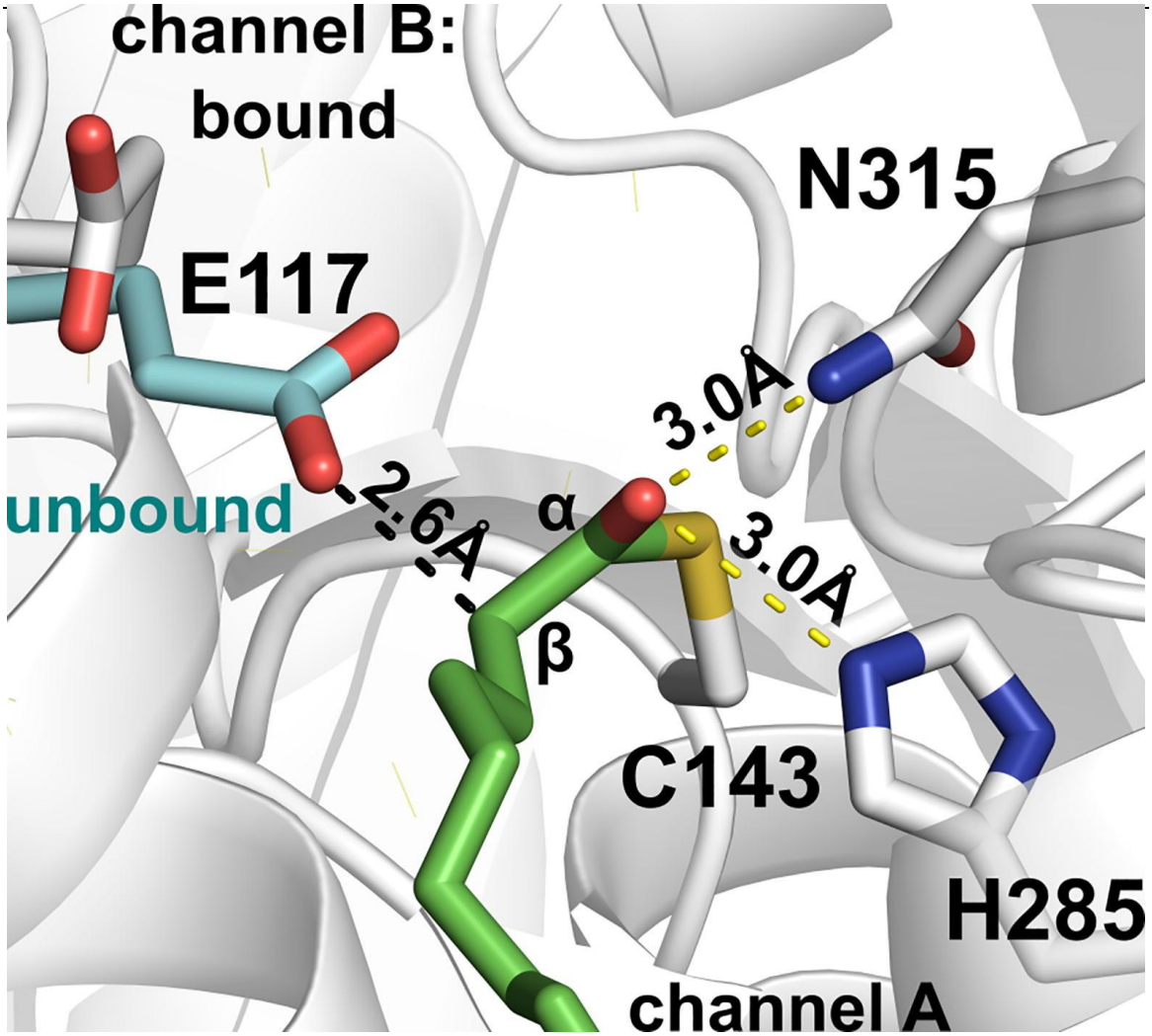


Figure 4-6. E117 position in the unbound form of OleA suggests its role in deprotonation of the A-Channel chain with stabilization of deprotonated β-methylene carbon by enolization facilitated by the oxyanion hole formed by H285 and N315.

Enzyme modeling to explain mechanism. In addition to offering a way to make novel β-keto acids, and hence β-lactones and lipids, the use of *p*-nitrophenyl esters to drive chain condensation provides new insights into the Claisen mechanism of OleA. Prior to this study, there was no direct evidence for the directionality of the Claisen condensation reaction. The use of *p*-nitrophenyl esters and ^{13}C-labeling allowed new insights to be obtained. OleA reaction-intermediate structures were modeled to better understand these new insights.

The positioning of E117 in OleA and its role in the directionality of the Claisen condensation reaction is re-examined here in light of these new findings. Thiolases demonstrated or predicted to be OleA enzymes show a [LIVYF]-E-P-X-X-[VIL] motif where the E corresponds to E117 in the *X. campestris* OleA studied here (Figure 4-6). E117 interposes into the active site catalytic triad of C143, H285 and N315 from the opposite subunit of the OleA homodimer. It has been proposed to be the catalytic base that abstracts a hydrogen atom to activate one of the chains. Consistent with that, a E117D mutation results in greatly diminished Claisen condensation activity, while mutation to an alanine completely eliminates this activity (34).

Overlay of the two subunits in the OleA structure PDB 4KU3, where channel B is bound with myristoyl-CoA or is unbound in either subunit, shows that residue E117 has multiple conformations. Modelling the thioester intermediate bound in Channel A and linked to the catalytic cysteine, C143, it appears that the E117 conformer, with Channel B unoccupied, is positioned well to act as a base that could deprotonate the methylene carbon of the thioester intermediate. When the acyl-CoA is bound in Channel B, the E117 conformer found in crystal structures is placed in a way that would least obstruct substrate entry or product release from either channel.

A previous X-ray crystallographic experiment with lauryl- and myristoyl-CoA revealed structures with a fatty acid bound in Channel A and acyl-CoA in Channel B (31). In those structures, PDB 4KU5 and 4KU3, the general base E117 (32, 34) was somewhat closer to the Channel B acyl-CoA methylene carbon than the corresponding Channel A fatty acid methylene carbon. This led

to an earlier proposal that the Claisen condensation might occur via the Channel B chain attack on the Channel A carbonyl carbon, opposite to what was demonstrated here. Note that the structures 4KU5 and 4KU3 do not represent reaction intermediates. The fatty acid chain is not tethered to C143 and an intervening oxygen molecule positions the methylene carbon hydrogen atoms considerably further away from the other reactive components. Moreover, other structures show a different positioning of E117. These alternative positions are depicted in Figure 4-6.

In active sites with Channel B not occupied with an acyl-coA, E117 is positioned closer to the catalytic residues and when modelling the enzyme thioester intermediate of the fatty acyl chain bound to C143 in these active sites, E117 is well positioned (2.6 Å) to abstract a proton from the methylene carbon of the fatty acyl chain (Figure 4-6). When the proton is abstracted, the deprotonated methylene carbon can be further stabilized in its enol form by the oxyanion hole formed by residues H285 and N315. Then, E117 can swing back to make room for the acyl-CoA to come into Channel B, positioning both chains for the subsequent Claisen reaction.

Conclusions

The present study provides new insights into the mechanism of, and applications for, the thiolase enzyme OleA. The acyl chains from *p*-nitrophenyl esters are transferred onto C143 and extend into Channel A, subsequently being activated to attack an acyl chain bound to Coenzyme A (CoA) in Channel B. Since *p*-nitrophenyl esters are much less expensive to purchase and can be more readily synthesized than CoA esters, these findings also provide a means to introduce novel functionality into OleA products. OleA may then be used with additional enzymes, OleBCD, to synthesize novel lipids and β-lactone natural products.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financially supported by National Institutes of Health Biotechnology training grant (5T32GM008347-27) and the University of Minnesota MnDRIVE program. We thank the U.S. Department of Energy Joint Genome Institute for synthetic DNA. The work conducted by the U.S. Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, is supported under [DE-AC02-05CH11231].

Author contributions

MDS and LPW conceived the study. MDS, LJT, TB, JER, LPW participated in research design. MDS, LJT, TB, JER performed the experiments. All authors analyzed data. MDS and LPW wrote the original draft of the manuscript and all authors approved the final version.

CHAPTER 5 – Generating CoA thioesters from *p*-nitrophenyl esters for use as enzyme substrates

"It does not matter how slowly you go as long as you do not stop."
— Confucius

Abstract

OleA is the first enzyme in the olefin pathway, responsible for the generation of industrially relevant olefinic hydrocarbons and β-lactone antibiotics. OleA has been shown to hydrolyze *p*-nitrophenyl esters through the first transesterification step of the OleA mechanism. Additionally, *p*-nitrophenyl esters have been shown to participate in the Claisen condensation of one *p*-nitrophenyl ester and one CoA thioester. This paper demonstrates how thiolysis of *p*-nitrophenyl esters by CoA can lead to the synthesis of CoA thioesters *in vitro*. OleA is able to use the resulting CoA thioester substrates, pulling the thiolysis reaction forward, and generating a β-keto acid originating from two *p*-nitrophenyl esters. Additionally, studies into improving the yield of condensation products revealed that the cationic detergent cetyltrimethylammonium bromide (CTAB) is able to increase β-keto acid yield by suppressing the rate of futile hydrolysis by OleA. We propose that CTAB stabilizes the OleA homodimer.

Introduction

Thiolase enzymes exist in all domains of life and catalyze a number of important reactions involved in rearrangement of carbon chains(117). A subsection of thiolase enzymes are able to carry out non-decarboxylative head-to-head Claisen condensation reaction using fatty acyl CoAs (Figure 5-1) (4, 6). One such enzyme is OleA. This is the first enzyme in a biosynthetic pathway capable of making either β-lactone compounds or olefinic hydrocarbons (12, 27, 100). Both of these compounds are important industrially as olefins can be used as potential biofuels, while β-lactones are useful as antiobesity, antitumor and

antibiotic medications (7, 12, 100, 101, 118, 119). OleA has been shown as a major determinant in the overall chemical structure of the β-lactone, making it an important enzyme for study (36, 43, 109).

Many enzymes beyond OleA use CoA. Coenzyme A is a widely used cofactor across the three domains of life. A search through the BRENDA database yields an estimated 8% of enzymes across all 7 EC number classifications that interact with CoA in some form (70), making this molecule highly important in biology. The CoA molecule itself owes much of its importance to its thiol group, which forms thioester bonds with various carboxylic acids. Thioesters have an advantage over oxygen esters in that they are 100x and 200x more reactive with amine and carbanion nucleophiles respectively yet have a similar stability in aqueous environments (47). The CoA thioester also imparts an increased acidity (pKa ~21) to the α-proton, facilitating deprotonation and formation of the enolate (36, 43). This enolate is able to carry out an attack of the carbonyl of another CoA thioester and form a carbon-carbon bond.

For all CoA-using enzymes, substrate expense and lack of structural diversity are barriers to mechanistic studies. Coenzyme A esters are not abundantly available commercially, are difficult to synthesize in high yield, and are among the most expensive of enzyme substrates (116). By contrast, commercial p-nitrophenyl esters are <1% the cost of corresponding CoA esters (120) and thousands of divergent analogs can be synthesized in > 85% yield from a carboxylic acid and p-nitrophenyl chloroformate via a convenient 2-hour synthetic coupling reaction (109).

In a previously published paper, we detailed how *p*-nitrophenyl hexanoate were used to screen for activity with 74 diverse OleA homologs (43). These OleA homologs were found to be able to hydrolyze the *p*-nitrophenyl ester, leading to a release of fatty acid and *p*-nitrophenol at rates similar to other esterases with similar substrates (43). In addition to the *p*-nitrophenyl hexanoate, we tested for activity with a panel of 14 other *p*-nitrophenyl esters with a wide range of chemical features (109). Our screening showed OleA hydrolysis activity with non-native structurally-divergent p-nitrophenyl esters (109). We hypothesized that

unique condensation products could be synthesized if Claisen condensation could be achieved starting with readily-synthesized *p*-nitrophenyl esters (Figure 5-1B). We were able to demonstrate that the *p*-nitrophenyl ester could replace one of the CoA thioester substrates and participate in the Claisen condensation reaction with a second CoA thioester to form the condensed β-keto acid product (35). This has implications for future biosynthesis of non-native yet industrially relevant β-lactone and hydrocarbon compounds (21).

While we could replace one fatty acyl-CoA substrate of OleA with a *p*-nitrophenyl ester, we were unable to replace the second CoA thioester with another *p*-nitrophenyl ester substrate (35). Here, we demonstrate one method to circumvent this limitation by promoting the acylation of CoA by *p*-nitrophenyl esters in order to generate novel CoA thioesters that then feed into the OleA Claisen condensation reaction (Figure 5-1C). Thus, we are able to produce novel β-keto acids with both arms originating from a *p*-nitrophenyl substrate.

One limitation for *in vitro* OleA activity is futile hydrolysis (29). We pursued the use of the ionic detergent CTAB as a way to promote thiolysis of *p*-nitrophenyl esters by CoA to increase acyl-CoA concentrations, leading to an increase in total β-keto acid yield. CTAB has been shown to greatly increase the rate of thiolysis by CoA when above the critical micelle concentration (CMC) (121–124). However, ionic detergents are known at concentrations above CMC to denature enzym e (125). We describe the effects of the application of CTAB below CMC in this paper anvd present evidence that while CTAB below CMC does not promote thiolysis it nonetheless improves overall OleA yield. We demonstrate how both yield and rate of condensation increases in the presence of 0.1 mM CTAB, while rate of futile hydrolysis is suppressed. We propose that CTAB may be involved in the stabilization of the OleA enzyme.

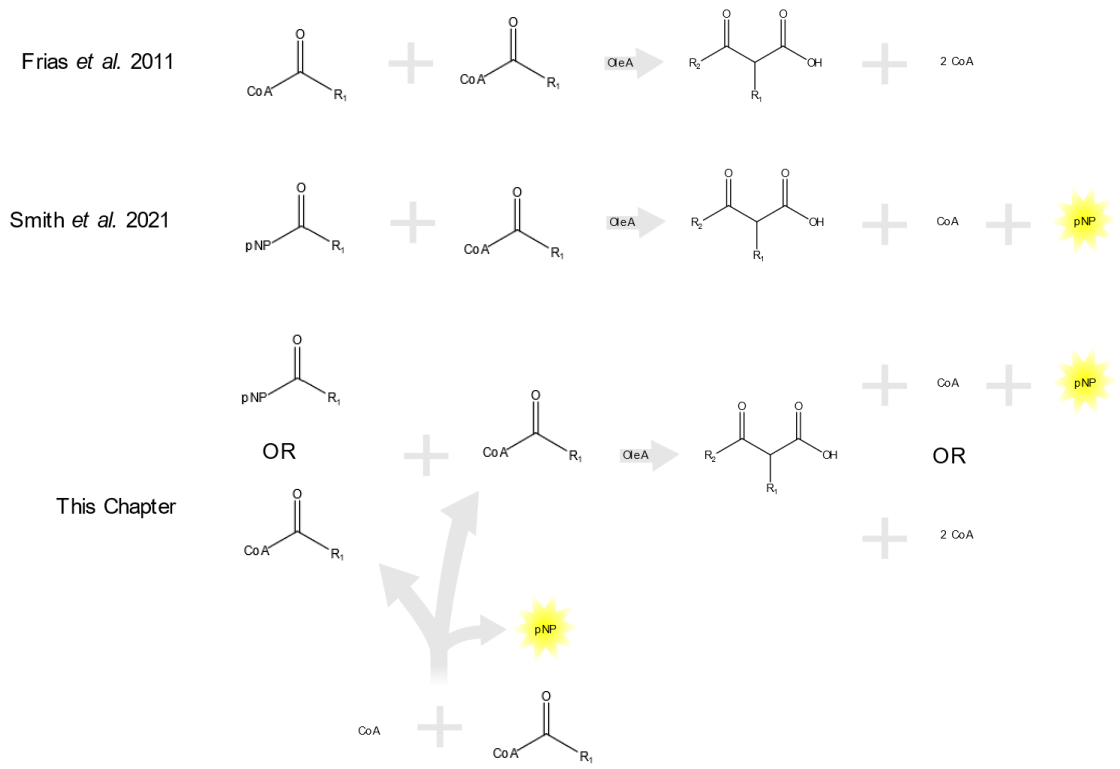


Figure 5-1. Schematics of the Condensation Reactions of OleA. A.) Schematic of two acyl-CoA thioesters reacting with OleA to create b-keto acid (35). B.) Schematic of one acyl-CoA thioester and one *p*-nitrohpenyl ester reacting with OleA to create b-keto acid (43). C.) Schematic of one *p*-nitrophenyl ester reacting with free CoA to create an acyl-CoA thioester, which then reacts with another *p*-nitrophenyl ester and OleA to create b-keto acid (this paper).

Materials and Methods

Chemicals and Reagents. *p*-Nitrophenyl acetate, decanoate, laurate, myristate, and hexanoate were purchased from Sigma Aldrich. Acetyl-CoA was purchased from CoALA Biosciences. SNAC, Decanoyl-CoA, lauryl-CoA, myristoyl-CoA, and hexanoyl-CoA were purchased from Sigma Aldrich. CoASH was purchased from [COMPANY NAME].

p-Nitrophenyl nonanoate and *p*-Nitrophenyl [NAME] were created according to the methods described in (109). Stock solutions were made at 4 mM dissolved in ethanol.

Protein Purification. OleA is expressed and purified according to the methods stated in Smith *et al* (43).

***p*-Nitrophenyl Acetate and SNAC Thiolysis Assay.** 100 µM *p*-nitrophenyl acetate was combined with 20 mM SNAC in a 50 mM TrisHCl buffer with a pH of 8.0. Total volume was 200 uL in a 96 well, F-bottom UV-STAR® MICROPLATE (Catalog: 655801). Absorbance was measured across a wavelength of 250-500 nm every 15 minutes for 75 minutes total at RT. The assay was run in a Epoch 2 Microplate Spectrophotometer (Biotek). Normalization was done by subtracting the absorbance detected from a well with only 20 mM SNAC in buffer.

***p*-Nitrophenyl Acetate and CoA Thiolysis Assay.** 100 µM *p*-nitrophenyl acetate was combined with 20 mM CoA and increasing amounts of CTAB from 0, 0.5, and 1 mM concentrations. The assay was run at 200 uL in a 96 well plate (catalog no. 25-104; Genesee Scientific) and read at 410 nm in a SpectraMax Plus384 microplate reader (Molecular Devices) colorimeter. Absorbance measurements were taken every minute for 75 minutes total.

Acyl-CoA Condensation Assay. 27 µg of purified OleA was incubated with 200 µM acyl-CoAs with carbon chain lengths 10, 12, 14, or 16 in a 500 µL reaction. Increasing amounts of CTAB at 0.01, 0.025, 0.05, 0.1, 0.25, 0.5 and 1 mM were added to the 50 mM TrisHCl pH 8.0 buffer. Every condition was repeated in triplicate and allowed to react overnight at RT. Extraction of β-keto acid was done by vortexing the reaction solution with 500 µL MTBE and removing the organic solvent. This was then run on an Agilent 7890a gas chromatograph and an Agilent5975c mass spectrometer (Santa Clara, CA). The method used was described previously in a 2021 paper by Smith *et al* (35). Both the heat from the GC-MS as well as the long incubation time results in the decarboxylation of the β-keto acid into ketone product.

Standards were prepared with increasing concentrations of commercially available 10-nonadecanone, 12-tricosanone, 14-heptacosanone, and 16-hentriacontanone in order to determine yield.

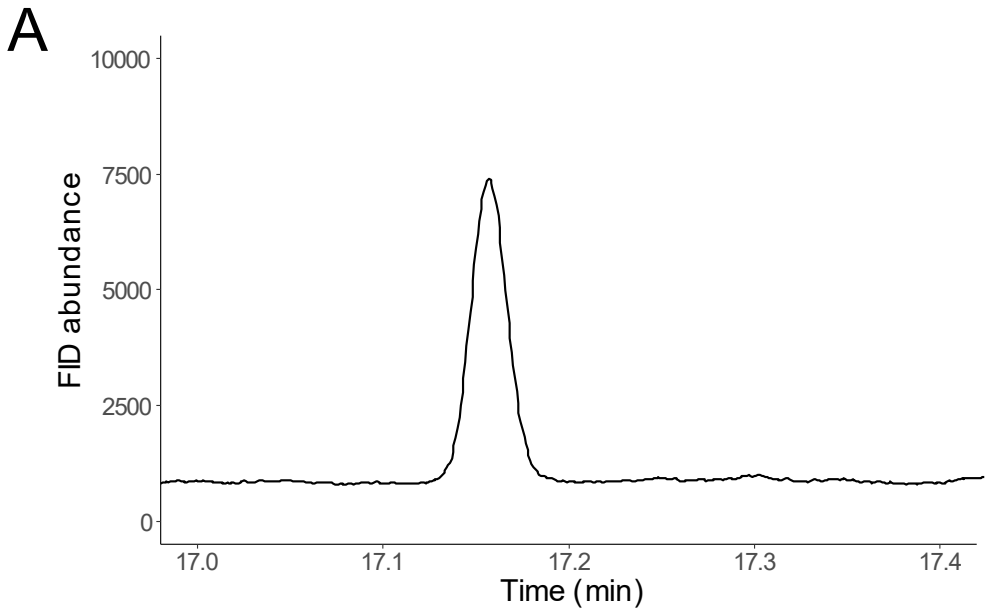
***p*-Nitrophenyl Condensation Assay.** 27 mg of purified OleA was incubated with 200 µM total *p*-nitrophenyl esters ranging in carbon chain lengths of 8, 9, 10, 12, 14, and 16 in a 500 µL reaction. This was done with 50 mM TrisHCl pH 8.0 buffer either with or without 0.1 mM CTAB. In one experiment the addition of OleA into a mixture containing the *p*-nitrphenyl ester and CoA was delayed by up to 60 minutes. Each condition was repeated in triplicate and allowed to react overnight at RT. Extraction of β-keto acid was done by vortexing the reaction solution with 400 µL of MTBE, plus 100 µL of diazomethane for methylation of the fatty acid. The solution was then run on GC-MS using the same method described previously.

Results and Discussion

Addition of CoA to reactions of OleA and *p*-Nitrophenyl esters forms product. We have previously shown that we are able to substitute *p*-nitrophenyl ester for one of the two acyl CoA required in a typical OleA condensation reaction (35), but we never observed condensation without the presence of an acyl-CoA. Even the substitution of one *p*-nitrophenyl ester for acyl-CoA opened the door for making non-native β-keto acids, however, the limitation of the requirement for acyl-CoA prevents complete control over the chemical structure as chemical synthesis of non-natural acyl-CoA substrates remain difficult and costly. We sought an alternative way to introduce variability in the second chain of the β-keto acid.

To determine whether CoA itself played a role in condensation, we incubated *p*-nitrophenyl esters with OleA along with CoA. In a previous paper, we showed how the first substrate transesterified to the active site cysteine was the substrate that was deprotonated by glu117 (35). However, we were still unable to use the *p*-nitrophenyl ester as the second substrate for Claisen condensation.

We hypothesized that CoA may be needed to show a conformational change but were skeptical that a combination of CoA and *p*-nitrophenyl ester would be enough to produce β-keto acid. To our surprise, the addition of free CoA into a reaction mixture containing OleA and *p*-nitrophenyl esters was all that was needed to observe the formation of β-keto acid. This was not the case when OleA was reacted with *p*-nitrophenyl esters and other thiols such as n-acetylcysteamine or SNAC. When SNAC were added to the reaction, we were unable to detect any β-keto acid (not shown). Thus, we propose that CoA thioesters are necessary as a second substrate in the Claisen condensation. This may be in order to properly position the second substrate in the alkyl channel B, or to induce a confirmation change. Early crystal structures of OleA with an acyl-CoA in the second substrate position have shown that while the pantetheine portion of the acyl-CoA does interact much with the pantetheine channel of the OleA active site, the adenine ring is located on the outside and indeed does interact with the surface of OleA (31, 32).


A
10000
7500
5000
2500
0
FID abundance
17.0
17.1
17.2
17.3
17.4
Time (min)

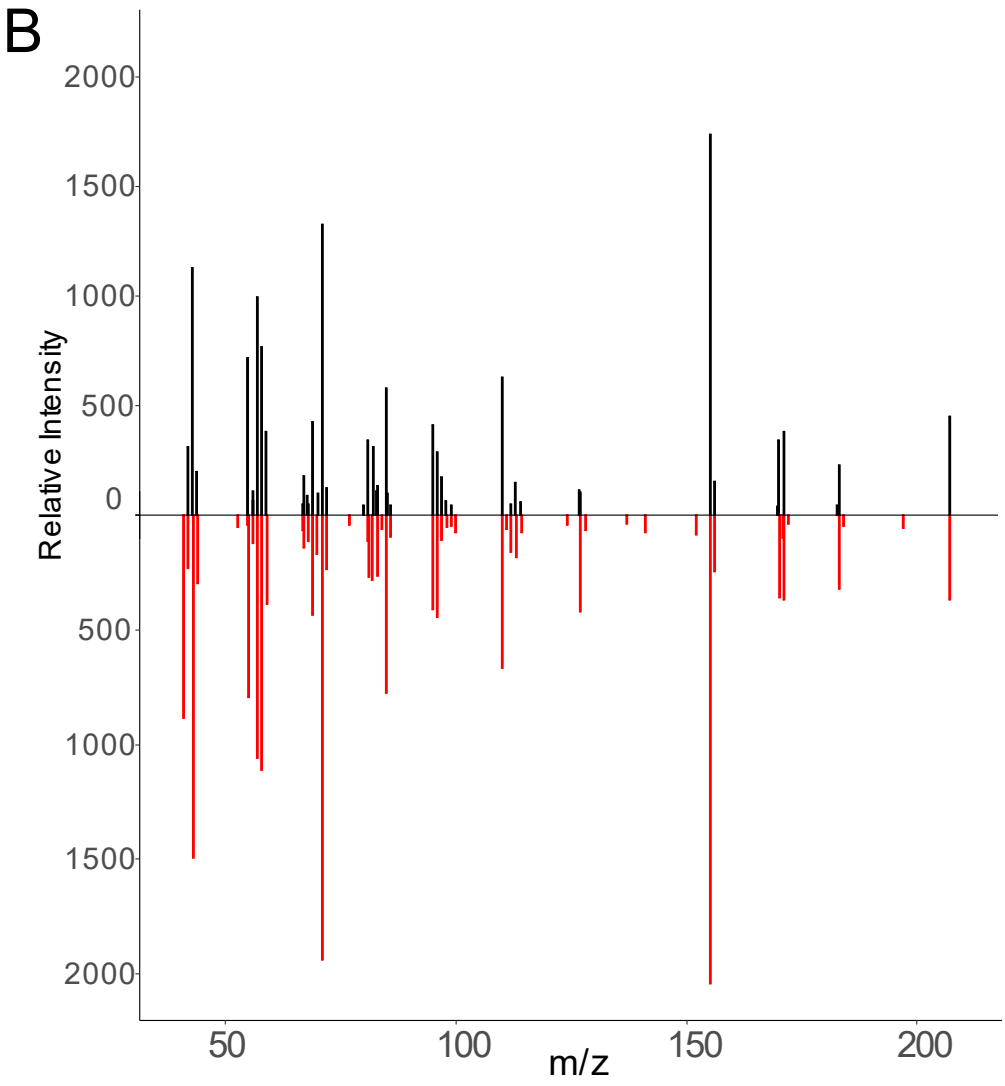

B
2000
1500
1000
500
0
500
1000
1500
2000
Relative Intensity
50
100
150
200
m/z

Figure 5-2. Formation of OleA condensation product. A.) GC-FID peak of 10-nonadecanone, which is the decarboxylated β-keto acid product formed by incubation of 27 µg OleA with 100 µM *p*-nitrophenyl dodecanoate, and 200 µM CoA overnight at RT. B.) MS spectra of the 10-nonadecanone peak in part A (above) compared to an MS spectra of a purchased 10-nonadecanone standard (below).

Acylation of CoA from *p*-nitrophenyl esters. We were interested in how CoA may be interacting with OleA and *p*-nitrophenyl esters to stimulate Claisen condensation while CoA simply induced a conformation change. In order to test this, we then incubated *p*-nitrophenyl esters with thiols such as free CoA or SNAC (N-acetylcysteamine) and noticed that the rate hydrolysis of *p*-nitrophenol increased relative to that in buffer. We hypothesized that thiolysis of the *p*-nitrophenyl ester by the free thiol was occurring, resulting in the transfer of the acyl chain from the *p*-nitrophenyl to the CoA (Figure 5-2A), as had been demonstrated in early studies of free thiols interacting with *p*-nitrophenyl esters (119–122). We followed the hydrolysis of 100 µM *p*-nitrophenyl acetate which was hydrolyzed by 20 mM SNAC until it reached equilibrium (Figure 5-2B, 5-2C). For this experiment, SNAC was chosen over CoA in order to reduce absorbance in the 250-300 range. This way we would be able to follow the reduction of absorbance due to the *p*-nitrophenyl ester which has an absorbance in this range. *p*-Nitrophenyl acetate was chosen as the ester due to its high solubility compared to longer chain lengths (53). This way solubility would not be a confounding factor. The accumulation of *p*-nitrophenol was observed at its absorbance of 410 nm. We then determined the equilibrium constant to be $k=2.03 \times 10^{-3}$ indicating that the equilibrium lies far to the left (Figure 5-2).

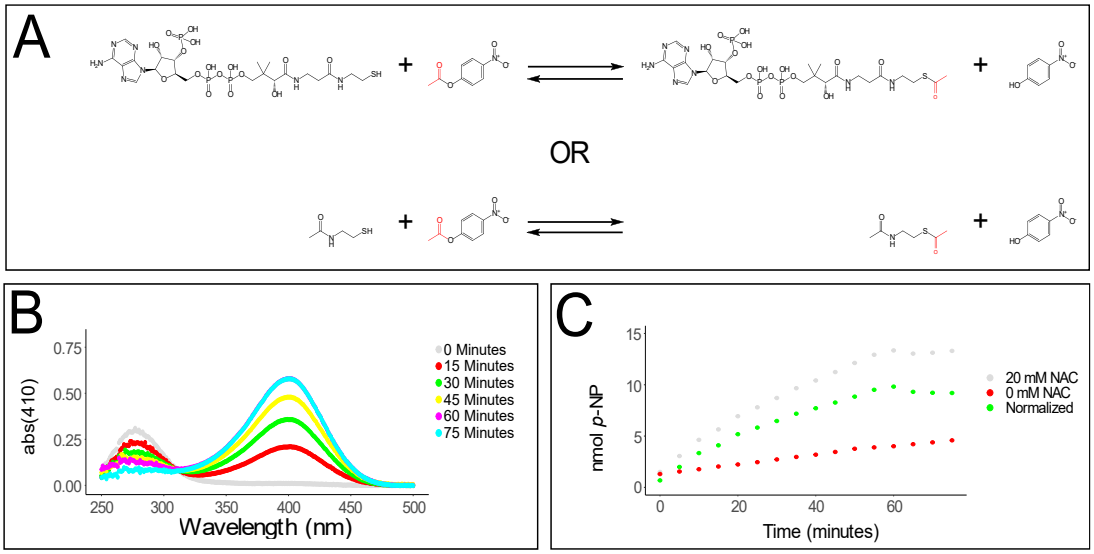


Figure 5-3. Acylation of SNAC and CoA following a thiolysis reaction with *p*-nitrophenyl acetate. A.) Schematic of thiolysis reaction with *p*-nitrophenyl acetate with either CoA or SNAC. B.) Graph of a reaction between 20 mM SNAC with 100 uM *p*-nitrophenyl acetate over a period of 75 minutes. Absorbance was recorded across a wavelength range of 250 to 500. C.) Graph of *p*-nitrophenol formation in n from either the reaction of SNAC and *p*-nitrophenyl acetate (grey), hydrolysis of *p*-nitrophenyl acetate in buffer (red), or the difference of the two graphs (green).

While we were able to show that thiolysis is occurring between SNAC and *p*-nitrophenyl acetate, and later confirmed with CoA (Figure 5-3), the equilibrium constant showed that the reaction would only result in a small amount of acyl-CoA present in solution. A small amount of acyl-CoA would not be expected to result in much if any detectable condensed OleA product. However, after incubating free CoA and *p*-nitrophenyl ester with OleA, we showed the formation of heat labile β-keto acid, seen as the more stable ketone in GC-MS (Figure 5-2) (29). Though OleA appeared to undergo futile hydrolysis more often than condensation when reacted with *p*-nitrophenyl esters and CoA (Table 5-1), there was no concentration of products needed after extraction. Instead extraction with

MTBE was done on a 1:1 scale. We hypothesize that OleA is able to pull the reaction forward (Figure 5-2A), forming more acyl-CoA thioester and ultimately forming a low but not insignificant amount of product.

	C8	C9	C10	C12	C14	C16
C8	0					
C9	0.5	1.8				
C10	1.3	3.7	3.8			
C12	1.1	6.1	4.2	4.4		
C14	0.6	2.9	3.1	3.6	0	
C16	0	0.8	1.6	2.7	0.7	0

Table 5-1. Percent condensation product over total reaction product of OleA reacted with CoA and *p*-nitrophenyl esters. 50 µM of two *p*-nitrophenyl ester chain lengths totaling 100 µM were reacted with 200 µM CoA and 27 µg OleA overnight at RT in 500 µL reaction mixtures. Product was then extracted and reacted with diazomethane to visualize fatty acid on the GC. FID peaks of either hydrolyzed fatty ester or condensed product were integrated, and the percentage of total condensed product over total reaction (hydrolyzed + condensed product) were tabulated above.

Of the *p*-nitrophenyl esters tested, any reaction mixtures that contained C8 and C16 chain length esters produced the lowest amounts of product. When first characterized, OleA had the lowest reactivity with octanoyl-CoA (C8), perhaps accounting for the low reactivity seen with the corresponding *p*-nitrophenyl octanoate (29). Though palmitoyl-CoA (C16) was found to be the best, the corresponding *p*-nitrophenyl palmitate has very low solubility, perhaps leading to a decrease in available ester to react with CoA prior to hydrolysis by buffer or OleA itself.

CTAB stimulates higher yield. We showed in the previous section that OleA could pull the thiolysis reaction forward in order to generate a β-keto acid. Still, this yield is low, and we were interested in exploring ways to enhance the yield.

Early chemical studies on thiolysis of *p*-nitrophenyl esters have shown that ionic surfactants such as CTAB (cetyl trimethyl ammonium bromide) can increase the rate of thiolysis (119–122). We reasoned that a stimulation of thiolysis could perhaps shift the reaction to the right in order to improve the yield of β-keto acid. Thus, we tested this through the addition of 0.1 and 0.5 mM CTAB to a reaction with 100 µM *p*-nitrophenyl acetate and 20 mM CoA (Figure 5-4A). The addition of CTAB at 0.5 mM was able to greatly increase the rate of thiolysis between CoA and *p*-nitrophenyl esters (Figure 5-4A), consistent with published research (119–122). While most published research on the stimulation of thiolysis of *p*-nitrophenyl esters with CTAB used concentrations of CTAB above critical micellular concentration (CMC), our studies focused on whether we saw the same effect below CMC. Notably we were only able to see a significant increase in rate of thiolysis once we approached the CMC of 0.95 mM (120, 122). By contrast, the rate increase of the thiolysis reaction was minimal at 0.1 mM CTAB (Figure 5-4A).

One reason for running reactions with CTAB below its CMC is because ionic detergents are known to denature enzymes (123). While CTAB stimulation of thiolysis of *p*-nitrophenyl esters by CoA was low at lower concentrations, it was hypothesized that 0.1 mM CTAB could still promote thiolysis if addition of OleA to the reaction was delayed. This would presumably allow acyl-CoA to build up in solution, allowing for higher condensation reaction compared to thiolysis. This way, OleA did not hydrolyze *p*-nitrophenyl esters before they could react with CoA. Astonishingly, the addition of 0.1 mM CTAB itself improved the yield nearly double what it was without CTAB (Figure 5-4B). The delay of OleA addition into solution had no effect whatsoever. One explanation for these results is that addition of CTAB does not play a role in thiolysis at these smaller concentrations, and therefore does not increase the pool of acyl-CoA to be used by OleA. Instead, it seemed that CTAB may play a role in stimulating OleA in some fashion. Therefore, we sought a mechanistic explanation for the positive effect of CTAB on the yield of the OleA reaction.

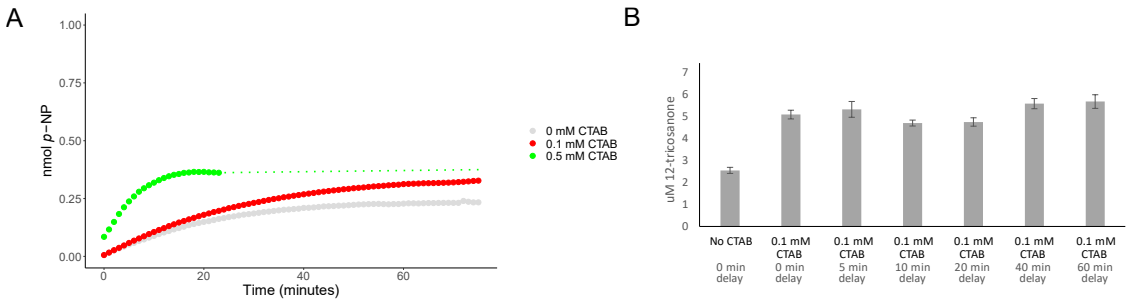


Figure 5-4. CTAB stimulation of thiolysis and condensation. A.) Thiolysis of 100 µM *p*-nitrophenyl acetate by 20 mM CoA in the presence of 0 mM (grey) 0.1 mM (red), and 0.5 mM (green) CTAB. Absorbance of the resulting *p*-nitrophenol was recorded at 410 nm in a 200 µL microtiter well plate reaction in a colorimeter. B.) Yield of β-keto acid measured as the decarboxylated ketone product by GC-MS. 500 µL reactions of 200 µM *p*-nitrophenyl dodecanoate with 400 µM CoA with or without 0.1 mM CTAB were run with 27 µg of OleA in triplicate. Addition of OleA to the reaction was delayed by either 0, 5, 10, 20, 40 or 60 minutes before allowing the reaction to proceed overnight.

CTAB stimulates OleA condensation. In order to figure out how CTAB was working to improve our yield, we incubated OleA and acyl-CoA thioesters with increasing amounts of CTAB (Figure 5-5A). Acyl-CoA thioesters were used instead of *p*-nitrophenyl esters with CoA to see whether CTAB stimulated yield in an enzymatic reaction that does not require the thiolysis reaction to create the acyl-CoA substrates. We found that increasing amounts of CTAB had a large increase in the yield of β-keto acid (Figure 5-5A). However, increasing past a certain concentration resulted in a decrease from the maximum yield observed where yield was lower than that with no CTAB (Figure 5-5A). The concentration of CTAB needed for the maximum β-keto acid yield was different for each chain length but varied between 0.05 and 0.25 mM. At these concentrations, the maximum yields increased up to 7-fold than what was seen when no CTAB was present. CTAB had a negligible effect on yield of 16-hentriacontanone. When testing the rate of condensation, 0.1 mM CTAB created an increase in rate over 3

times the original rate without CTAB (Figure 5-5B). However, when looking solely at rate of hydrolysis with *p*-nitrophenyl decanoate, we saw a depression in the rate of 96% (Table 5-2). Thus, we believe that CTAB works to stimulate condensation by reducing the amount of futile hydrolysis.

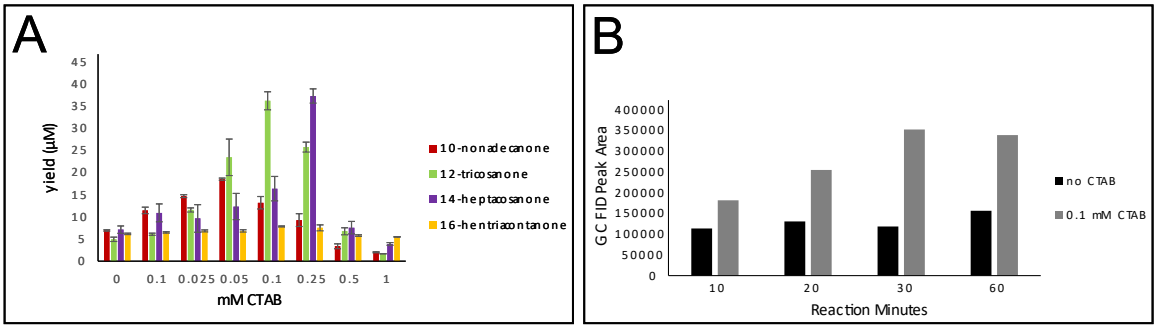


Figure 5-5. CTAB stimulates increased yield from OleA condensation reaction. A.) Yield of 10- nonadecanone, 12-tricosanone, 14-heptacosanone, and 16-hentriacontanone. 0, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, and 1 mM CTAB. 27 µg of OleA was incubated with 200 µM acyl-CoA with chain lengths of 10, 12, 14, or 16 and reacted in 500 µL of 50 mM TrisHCl pH 8 overnight at RT. µM amounts of ketone were determined through standard curves of commercially purchased standards for each ketone present (Supplemental figures 5-S# through 5-S#). B.) Stimulation in rate of condensation product seen in the presence of 0.1 mM CTAB. Dodecanoyl-CoA was reacted with OleA in the presence and absence of 0.1 mM CTAB for the duration of 1 hour. The integration of GC-FID peaks are shown above demonstrating an increase in rate of condensation product formation over the span of the hour.

Condition	Rate (nmol p-NP/min)	% Decrease in Hydrolysis Activity
p-NP acetate	0.31 ± 0.2	
p-NP acetate + CTAB	0.075 ± 0.03	75
p-NP acetate + OTAB	0.14 ± 0.05	53
p-NP hexanoate	4.8 ± 1	
p-NP hexanoate + CTAB	0.23 ± 03	95
p-NP hexanoate + OTAB	0.39 ± 0.16	92
p-NP laurate	1.6 ± 0.4	
p-NP laurate + CTAB	0.061 ± 0.01	96
p-NP laurate + OTAB	0.22 ± 0.09	86

Table 5-2. Table of rates of hydrolysis for 100 µM *p*-nitrohpenyl acetate, hexanoate, and dodecanoate reacted with 27 µg OleA in the presence and absence of 0.1 mM CTAB or OTAB. 200 µL reactions are followed by absorbance of released *p*-nitrophenol at 410 nm for 30 minutes. Rates are then calculated and tabulated above.

Currently it is unknown why CTAB can suppress futile hydrolysis for OleA. Though CTAB is below CMC in these experiments, it is known that small aggregates of CTAB can still form (124). These could interact with aggregates of long chain *p*-nitrophenyl esters that are known to form in aqueous solution (53). In some cases, aggregates of CTAB are known to be catalytically active and have been known to cause Claisen rearrangement (125), though when tested for Claisen condensation in a reaction containing *p*-nitrophenyl and CoA without the presence of OleA, no condensation product was observed (not shown).

An enzyme known as Xanthine oxidase was shown to interact with CTAB yet low concentrations of CTAB led to an inhibition of activity (126). A paper exists that details how CTAB improved the hydrolysis of α-Chymotripsyn in aqueous buffer (52, 127). Though here, CTAB, among other cationic surfactants were used at or above CMC (1 mM) with rate of activity of α-Chymotripsin increasing over CMC before finally decreasing at 5 mM (127). Thus, none of these examples quite explain the results we see in this chapter. We propose that

CTAB may be stabilizing the protein. CTAB is also known to interact with proteins. CTAB is known to solubilize recombinant protein from inclusion bodies for purification (128). In other experiments, CTAB in low concentrations has been shown to stabilize enzymes, acting as an ‘artificial chaperone’ in refolding heat-degraded lysozyme (129). CTAB below CMC has been shown to interact with protein and has been posited that it could help stabilize membrane proteins especially(130). OleA, and other Ole proteins are thought to interact with the membrane (92), and thus may benefit from the addition of sub-CMC CTAB. Further study is needed in order to determine how CTAB interacts with the protein and the true mechanism by which it improves the Claisen condensation of OleA.

CHAPTER 6 – Conclusions

"The thing about finishing a story is that finishing is really only the beginning."– William Herring

Conclusions

This work details a significant improvement in the understanding of the thiolase enzyme OleA, in both its diversity and mechanism. It is a first step in the biosynthesis of novel β-lactones and related natural products.

In chapter 2, a colorimetric assay was developed that acts as a rapid and accurate screen for the identification of OleA (43). This screen helps to save time in accurately choosing putative OleA homologs for purification and further study. This screen was expanded to be able to test for multiple substrates due to the relative ease of synthesizing *p*-nitrophenyl esters from a carboxylic acid of choice. Older methods relied on purified protein, and the purification of each OleA homolog could be lengthy. With this assay, the time for identification of putative OleAs was greatly reduced.

Because the pNP ester colorimetric assay allowed for easy testing of multiple substrates with multiple OleAs, large amounts of data can be generated that are useful for bioinformatic analysis. The data presented in figure 3-2 represents 1095 distinct points of data and were generated in only two months' time(109). Chapter 3 details how the data was used by a skilled bioinformatician to develop a machine-learning model that could predict substrate specificity for novel OleAs. This work has implications beyond OleA itself, as it is proof-of-principle for how this method, which uses both enzyme and substrate features to predict enzyme substrate specificity can be applied more generally to other biosynthetic enzymes. The natural expansion of this research involves using the predictions to help design OleAs or other biosynthetic enzymes with desired activities. Thus, this assay can have far reaching applications both for OleA and other enzymes.

In chapter 4, we detailed how the *p*-nitrophenyl ester could only participate as the first substrate in the OleA mechanism and undergo Claisen condensation with an acyl-CoA thioester acting as the second substrate. This made it possible to distinguish between two substrates to further define the mechanism. The *p*-nitrophenyl ester, an OleA-substrate analog allowed for the determination of the directionality of the Claisen condensation, an experiment that had not been possible when using acyl-CoA thioesters as the sole substrates. This is due to the fact that CoA thioesters can act as either substrate 1 or 2, making it impossible to determine which of the two acyl chains were deprotonated at the α-carbon position. In fact, earlier research looking at crystal structures of OleA with bound substrate suggested that the acyl chain that was deprotonated originated from the second substrate, as Glu117 was determined to be too far to act as the base for the alkyl chain of the first substrate. However, my research using pNP ester as the first substrate has shown that it is the first substrate that is deprotonated. With this new information we re-analyzed the available crystal structures of OleA and observed that Glu117 is found in a variety of positions in different crystal structures. It was through this research that we developed the hypothesis that Glu117 can swing between both positions in order to facilitate the deprotonation and later to allow for the Claisen condensation to take place. Thiolases are widespread among all domains of life, as are enzymes that utilize CoA thioesters and this method may be useful in the mechanistic understanding of enzymes beyond OleA (47, 116).

Finally, in chapter 5, I demonstrated the biotechnological utility of the *p*-nitrophenyl esters. As stated in earlier chapters, *p*-nitrophenyl esters are less expensive, easier to synthesize, and have multiple structures available commercially, while acyl-CoA thioesterases are mostly available as the natural substrates found in cells and have a much larger cost. So, it became an interesting avenue of research to investigate whether these esters could be used to generate β-keto acid, which could presumably be used to make industrially relevant β-lactones. Surprisingly, we found that not only could *p*-nitrophenyl esters be combined with CoA thioesters to change one of the sides of the β-keto

acid, but the thiolytic ability of CoA could be used to generate CoA thioesters from *p-nitrophenyl* esters for use with OleA. While this reaction alone resulted in a limited amount of CoA thioester product, we found that OleA could use those products as substrates and could therefore pull the initial thiolysis reaction forward. Ultimately this resulted in the creation of a β-keto acid where both side chains originated from *p*-nitrophenyl substrates. Finally, we show how CTAB, a cationic detergent, improves the yield of β-keto acid, both when reacting with CoA thioesters and *p*-nitrophenyl esters. We also show that CTAB not only stimulates the rate and total yield of product, but that it may do this by reducing futile hydrolysis by OleA. This has the potential to be useful in the use of other enzymes for biotechnology.

Future Directions

Characterizing Novel OleA Homologs

This research made use of 72 putative OleA homologs for the development of a colorimetric screen. However, many of the OleA homologs had interesting activity with a variety of *p*-nitrophenyl esters. *Kytococcus sedentarius* had the largest activity profile, and *Actinoplanes atraurantic*us had only 29.1% sequence similarity to OleA yet was active with many *p*-nitrophenyl esters tested. Further research can be done to understand the difference between the different homologs.

Another homolog of interest is NltAB. These enzymes were purified in collaboration with Dr. Serina Robinson (39) and represent an example of a heterodimeric OleA from a pathway where the final product is a β-lactone similar to the LstAB from the lipstatin pathway(40). From this, multiple research questions arise. What differences exist between the heterodimeric and homodimeric thiolases? Does this confer any ability for the heterodimer to avoid inactivation by the final product β-lactone, and if not, how does the cell prevent this? Does a complex exist for other pathways as one does for the Ole pathway, and if so, do pathways containing a heterodimer participate transiently with the complex similarly to OleA?

The generation of multiple OleA homologs from our collaboration with the Joint Genome Institute has provided an avenue of research that could be capitalized on by other researchers to further understand the diversity of OleA thiolases across microorganisms.

OleA Engineering

The machine-learning prediction method developed for substrate specificity gave rise to novel targets for OleA engineering. Mutations can be made to test *in vivo* whether these will result in a difference in substrate specificity. Additionally, while the screen tested multiple *p*-nitrophenyl ester substrates for a variety of OleAs, the substrate specificity model may only hold true for the first channels. This is because the *p*-nitrophenyl esters only react in the first step of the mechanism and thus only occupy channel A. Some β-lactone natural products such as lipstatin and ebelactone A have a difference in the two chain lengths of the b-lactone structure (6, 12, 131), and so it can be concluded that the channels of some OleA homologs indeed have different substrate specificity. Research should be done to look specifically at channel B. Additionally, it would be interesting to see if LstAB is truly only able to accept specific chain lengths into its active site that vary based on the channel, and whether we could engineer a homolog to do the same.

Generating Novel β-Lactones

While we were able to show that CTAB could improve the yield of OleA, more research must be done into the mechanism of how this works. By understanding this, we may be able to further improve OleA yield. Additionally, while CTAB stimulated OleA, we do not know whether CTAB would have a beneficial or deleterious effect on the other enzymes in the pathway.

In general, combining the research described in this thesis with the rest of the pathway remains an unexplored avenue of research. Introducing other enzymes and their cofactors can lead to unexpected consequences in the production of natural products *in vitro* (25, 26). It may be that OleA is able to

generate β-keto acids with side groups other than the naturally occurring acyl chains, but these may not be accepted by the downstream enzymes. Additionally, the inhibition of OleA by β-lactones described in chapter 1 presents another hurdle to overcome for production of β-lactones. It is pertinent to determine the mechanism of inhibition, as well as how to guard against it in future engineering of β-lactone synthesis. The work done in this thesis can be expanded upon and used to create β-lactones, or other natural products. Through this, it is my hope that this work could be leveraged to develop new therapeutics that are easier to produce and less expensive for patients that could benefit from them.

References

1. Wang Y, Tennyson RL, Romo D. 2004. β-Lactones: Intermediates for natural product total synthesis and new transformations. Heterocycles 64:605–658.

2. Getzler YDYL, Mahadevan V, Lobkovsky EB, Coates GW. 2002. Synthesis of β-lactones: A highly active and selective catalyst for epoxide carbonylation. Journal of the American Chemical Society 124:1174–1175.

3. Robinson SL, Wackett LP. 2020. Rings of Power: Enzymatic Routes to β-Lactones, p. 323–345. *In* Comprehensive Natural Products III. Elsevier.

4. Robinson SL, Christenson JK, Wackett LP. 2019. Biosynthesis and chemical diversity of β-lactone natural products. Natural Product Reports. Royal Society of Chemistry 36(3): 458-475.

5. Zhu T, Wang L, Wang W, Hu Z, Yu M, Wang K, Cui Z. 2014. Enhanced production of lipstatin from Streptomyces toxytricini by optimizing fermentation conditions and medium. Journal of General and Applied Microbiology 60:106–111.

6. Christenson JK, Richman JE, Jensen MR, Neufeld JY, Wilmot CM, Wackett LP. 2017. β-Lactone synthetase found in the olefin biosynthesis pathway. Biochemistry 56:348–351.

7. European Organisation for Research and Treatment of Cancer - EORTC, Celgene, Canadian Cancer Trials Group. A Phase III Trial of With Marizomib in Patients With Newly Diagnosed Glioblastoma (MIRAGE). NCT03345095.

8. Richardson RD, Ma G, Oyola Y, Zancanella M, Knowles LM, Cieplak P, Romo D, Smith JW. 2008. Synthesis of novel β-lactone inhibitors of fatty acid synthase. Journal of Medicinal Chemistry 51:5285–5296.

9. Zhu M, Harshbarger WD, Robles O, Krysiak J, Hull KG, Cho SW, Richardson RD, Yang Y, Garcia A, Spiegelman L, Ramirez B, Wilson CT, Yau JA, Moore JT, Walker CB, Sacchettini JC, Liu WR, Sieber SA, Smith JW, Romo D. 2017. A strategy for dual inhibition of the proteasome and fatty acid synthase with belactosin C-orlistat hybrids. Bioorganic & Medicinal Chemistry 25(11):2901-2916.

10. Böttcher T, Sieber SA. 2009. Structurally refined β-lactones as potent inhibitors of devastating bacterial virulence factors. ChemBioChem 10:663–666.

11. Shirley M. 2014. β-Lactones as key building blocks: Synthetic aplications to diverse natural products. Doctoral dissertation, Texas A & M University. Available electronically from https://hdl.handle.net/1969.1/152839.

12. Bai T, Zhang D, Lin S, Long Q, Wang Y, Ou H, Kang Q, Deng Z, Liu W, Tao M. 2014. Operon for biosynthesis of lipstatin, the beta-lactone inhibitor of human pancreatic lipase. Applied and Environmental Microbiology 80:7473–7483.

13. Kong K-F, Schneper L, Mathee K. 2010. Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. APMIS 118(1):1-36.

14. Fernandes R, Amador P, Prudêncio C. 2013. β-Lactams: chemical structure, mode of action and mechanisms of resistance. Reviews in Medical Microbiology 24(1):7-17.

15. Frias JA, Richman JE, Wackett LP. 2009. C29 Olefinic Hydrocarbons Biosynthesized by Arthrobacter Species. Applied and Environmental Microbiology 75(6):1774-1777.

16. Beller HR, Goh E-B, Keasling JD. 2010. Genes Involved in Long-Chain Alkene Biosynthesis in *Micrococcus luteus*. Applied and Environmental Microbiology 76(4):1212-1223.

17. Nichols DS, Nichols PD, McMeekin TA. 1995. A new n-C 31:9 polyene hydrocarbon from Antarctic bacteria. FEMS Microbiology Letters 2(3):281-285.

18. Chen HP, Zhao ZZ, Li ZH, Dong ZJ, Wei K, Bai X, Zhang L, Wen CN, Feng T, Liu JK. 2016. Novel Natural Oximes and Oxime Esters with a Vibralactone Backbone from the Basidiomycete Boreostereum vibrans. ChemistryOpen 5(2):142–149.

19. Zhao P-J, Yang Y-L, Du L, Liu J-K, Zeng Y. 2013. Elucidating the Biosynthetic Pathway for Vibralactone: A Pancreatic Lipase Inhibitor with a Fused Bicyclic β-Lactone. Angewandte Chemie International Edition 52(8)2298-2302.

20. Eustaquio AS, McGlinchey RP, Liu Y, Hazzard C, Beer LL, Florova G, Alhamadsheh MM, Lechner A, Kale AJ, Kobayashi Y, Reynolds KA, Moore BS. 2009. Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. Proceedings of the National Academy of Sciences 106(30) 12295-12300.

21. Schaffer JE, Reck MR, Prasad NK, Wencewicz TA. 2017. β-Lactone formation during product release from a nonribosomal peptide synthetase. Nature Chemical Biology 13:737-744.

22. Scherlach K, Hertweck C. 2009. Triggering cryptic natural product biosynthesis in microorganisms. Organic and Biomolecular Chemistry 7(9):1753–1760.

23. Yang HW, Romo D. 1999. Tetrahedron report number 492 Methods for the Synthesis of Optically Active [3-Lactones (2-Oxetanones). Tetrohedron 55(21):6403–6434.

24. Yang HW, Romo D. 1997. A Highly Diastereoselective, Tandem Mukaiyama Aldol-Lactonization Route to-Lactones: Application to a Concise Synthesis of the Potent Pancreatic Lipase Inhibitor, (-)-Panclicin D. Journal of Organic Chemistry 62(1):4–5.

25. Taniguchi H, Okano K, Honda K. 2017. Modules for in vitro metabolic engineering: Pathway assembly for bio-based production of value-added chemicals. Synthetic and Systems Biotechnology 2(2):65-74.

26. Lowry B, Walsh C, Khosla C. 2015. *In Vitro* Reconstitution of Metabolic Pathways: Insights into Nature's Chemical Logic. Synlett 26(8):1008-1025.

27. Christenson JK, Robinson SL, Engel TA, Richman JE, Kim AN, Wackett LP. 2017. OleB from Bacterial Hydrocarbon Biosynthesis Is a β-Lactone Decarboxylase That Shares Key Features with Haloalkane Dehalogenases. Biochemistry 56(40):5278–5287.

28. Christenson JK, Jensen MR, Goblirsch BR, Mohamed F, Zhang W, Wilmot CM, Wackett LP. 2017. Active multienzyme assemblies for long-chain olefinic hydrocarbon biosynthesis. Journal of Bacteriology 199(9)e00890.

29. Frias JA, Richman JE, Erickson JS, Wackett LP. 2011. Purification and characterization of OleA from Xanthomonas campestris and demonstration of a non-decarboxylative claisen condensation reaction. Journal of Biological Chemistry 286(13):10930–10938.

30. Bonnett SA, Papireddy K, Higgins S, del Cardayre S, Reynolds KA. 2011. Functional Characterization of an NADPH Dependent 2-Alkyl-3-ketoalkanoic Acid Reductase Involved in Olefin Biosynthesis in *Stenotrophomonas maltophilia*. Biochemistry 50(44):9633-9640.

31. Goblirsch BR, Jensen MR, Mohamed FA, Wackett LP, Wilmot CM. 2016. Substrate trapping in crystals of the thiolase olea identifies three channels that enable long chain olefin biosynthesis. Journal of Biological Chemistry 291(52):26698–26706.

32. Goblirsch BR, Frias JA, Wackett LP, Wilmot CM. 2012. Crystal structures of xanthomonas campestris OleA reveal features that promote head-to-head condensation of two long-chain fatty acids. Biochemistry 51(20):4138–4146.

33. Jensen MR, Goblirsch BR, Esler MA, Christenson JK, Mohamed FA, Wackett LP, Wilmot CM. 2018. The role of OleA His285 in orchestration of long-chain acyl-coenzyme A substrates. FEBS Letters 592(6):987–998.

34. Jensen MR, Goblirsch BR, Christenson JK, Esler MA, Mohamed FA, Wackett LP, Wilmot CM. 2017. OleA Glu117 is key to condensation of two fatty-acyl coenzyme A substrates in long-chain olefin biosynthesis. Biochemical Journal 474(23):3871–3886.

35. Smith MD, Tassoulas LJ, Biernath TA, Richman JE, Aukema KG, Wackett LP. 2021. *p*-Nitrophenyl esters provide new insights and applications for the thiolase enzyme OleA. Computational and Structural Biotechnology Journal 19:3087–3096.

36. Sukovich DJ, Seffernick JL, Richman JE, Gralnick JA, Wackett LP. 2010. Widespread head-to-head hydrocarbon biosynthesis in bacteria and role of OleA. Applied and Environmental Microbiology 76(12):3850–3862.

37. Knoot CJ, Pakrasi HB. 2019. Diverse hydrocarbon biosynthetic enzymes can substitute for olefin synthase in the cyanobacterium *Synechococcus sp.* PCC 7002. Scientific Reports 9:1360.

38. Robinson SL, Christenson JK, Richman JE, Jenkins DJ, Neres J, Fonseca DR, Aldrich CC, Wackett LP. 2019. Mechanism of a Standalone β-Lactone Synthetase: New Continuous Assay for a Widespread ANL Superfamily Enzyme. ChemBioChem 20(13):1701–1711.

39. Robinson SL, Terlouw BR, Smith MD, Pidot SJ, Stinear TP, Medema MH, Wackett LP. 2020. Global analysis of adenylate-forming enzymes reveals b-lactone biosynthesis pathway in pathogenic nocardia. Journal of Biological Chemistry 295(44):14826–14839.

40. Zhang D, Zhang F, Liu W. 2019. A KAS-III Heterodimer in Lipstatin Biosynthesis Nondecarboxylatively Condenses C8 and C14 Fatty Acyl-CoA Substrates by a Variable Mechanism during the Establishment of a C22 Aliphatic Skeleton. Journal of the American Chemical Society 141(9):3993–4001.

41. Mikami Y, Yazawa Y, Tanaka Y, Ritzau M, Gräfeb U. 1999. Isolation and Structure of Nocardiolactone, A New Dialkyl-Substituted β-Lactone From Pathogenic Nocardia Strains. Natural Product Letters 13(4)277-284.

42. Sun Y-P, Wu Y. 2005. An Expeditious Synthesis of Nocardiolactone. Synlett. 9:1477-1479

43. Smith MD, Robinson SL, Molomjamts M, Wackett LP. 2020. *In vivo* assay reveals microbial oleA thiolases initiating hydrocarbon and β-lactone biosynthesis. mBio 11(2):e00111-00120.

44. Steen EJ, Kang Y, Bokinsky G, Hu Z, Schirmer A, McClure A, del Cardayre SB, Keasling JD. 2010. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. Nature 463:559–562.

45. Tan GY, Liu T. 2017. Rational synthetic pathway refactoring of natural products biosynthesis in actinobacteria. Metabolic Engineering. 39:228-236.

46. Parsons JB, Rock CO. 2013. Bacterial lipids: Metabolism and membrane homeostasis. Progress in Lipid Research 52(3):249-276.

47. Haapalainen AM, Meriläinen G, Wierenga RK. 2006. The thiolase superfamily: Condensing enzymes with diverse reaction specificities. Trends in Biochemical Sciences 31(1):64-71

48. Bonk BM, Tarasova Y, Hicks MA, Tidor B, Prather KLJ. 2018. Rational design of thiolase substrate specificity for metabolic engineering applications. Biotechnology and Bioengineering 115(9):2167–2182.

49. Ammon R, Zoch E. 1958. The esterase activity of commercial papain (In German). Hoppe-Seyler´s Zeitschrift für physiologische Chemie 313.

50. Gupta R, Rathi P, Gupta N, Bradoo S. 2003. Lipase assays for conventional and molecular screening: an overview. Biotechnology and Applied Biochemistry 37(1):63-71.

51. Zha D, Xu L, Zhang H, Yan Y. 2014. Molecular Identification of Lipase LipA from *Pseudomonas protegens* Pf-5 and Characterization of Two Whole-Cell Biocatalysts Pf-5 and Top10lipA. Journal of Microbiology and Biotechnology 24(5):619-628.

52. Mcdonald CE, Balls AK. 1956. Transesterification reactions catalyzed by chymotrypsin. Journal of Biological Chemistry 221(2):993–1003.

53. Guthrie JP. 1973. Aggregation of *p*-Nitrophenyl Alkanoates in Aqueous Solution. Limitations on Their Use as Substrates in Enzyme Model Studies. Canadian Journal of Chemistry 51(21):3494-3498.

54. Burdette RA, Quinns DM. 1966. Interfacial Reaction Dynamics and Acyl-enzyme Mechanism for Lipoprotein Lipase-catalyzed Hydrolysis of Lipid *p*-Nitrophenyl Esters. The Journal of Biological Chemistry 261(26):12016–12021.

55. Schwaneberg U, Otey C, Cirino PC, Farinas E, Arnold FH. 2001. Cost-Effective Whole-Cell Assay for Laboratory Evolution of Hydroxylases in *Escherichia coli*. Journal of Biomolecular Screening 6(2):111-117.

56. Gadosy TA, Boyd MJ, Tee OS. 2000. Catalysis of Ester Aminolysis by Cyclodextrins. The Reaction of Alkylamines with *p* -Nitrophenyl Alkanoates . The Journal of Organic Chemistry 65(21):6879-6889.

57. Sims D, Brettin T, Detter JC, Han C, Lapidus A, Copeland A, del Rio TG, Nolan M, Chen F, Lucas S, Tice H, Cheng J-F, Bruce D, Goodwin L, Pitluck S, Ovchin-nikova G, Pati A, Ivanova N, Mavrommatis K, Chen A, Palaniappan K, D'haeseleer P, Chain P, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Schneider S, Göker M, Pukall R, Kyrpides NC, Klenk H-P. 2009. Complete genome sequence of *Kytococcus sedentarius* type strain (541T). Standards in Genomic Sciences 1(1):12-20.

58. Mathlouthi N, Traore SI, Cimmino T, Khelaifia S, Nguyen TT, Cadoret F, Couderc C, Raoult D, Rolain J-M. 2017. Genome sequence and description of *Mobilicoccus massiliensis sp*. *nov*. isolated from the stool of a Nigerian boy with kwashiorkor. New Microbes and New Infections 20:18-21.

59. Kang I, Lim Y, Cho J-C. 2018. Complete genome sequence of *Granulosicoccus antarcticus* type strain IMCC3135T, a marine gammaproteobacterium with a putative dimethylsulfoniopropionate demethylase gene. Marine Genomics 37:176-181.

60. Zhang Y, Zhang J, Fan L, Pang H, Xin Y, Zhang X. 2012. *Actinoplanes atraurantiacus* sp. nov., isolated from soil. International Journal of Systematic and Evolutionary Microbiology 62(10):2533-2537.

61. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk H-P, Clément C, Ouhdouch Y, van Wezel GP. 2016. Taxonomy, Physiology, and Natural Products of Actinobacteria. Microbiology and Molecular Biology Reviews 80(1):1-43.

62. TSUKAMURA M, MIZUNO S. 1971. *Mycobacterium obuense*, a Rapidly Growing Scotochromogenic *Mycobacterium* Capable of Forming a Black Product from p-Aminosalicylate and Salicylate. Journal of General Microbiology 68(2):129-134.

63. Koval SF, Williams HN, Stine OC. 2015. Reclassification of *Bacteriovorax marinus* as *Halobacteriovorax marinus gen. nov*., comb. *nov*. and *Bacteriovorax litoralis* as *Halobacteriovorax litoralis* comb. *nov*.; description of *Halobacteriovoraceae* fam. nov. in the class Deltaproteobacteria. International Journal of Systematic and Evolutionary Microbiology 65(2):593-597.

64. Biggs AI. 1954. A spectrophotometric determination of the dissociation constants of *p*-nitrophenol and papaverine. Transactions of the Faraday Society 50:800-802.

65. di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, Taly J-F, Notredame C. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Research 39(2): W13-W17.

66. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7(10):e1002195.

67. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai H-C, Zakai UI, Mitchell DA. 2017. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nature Chemical Biology 13:470-478.

68. Wright ES. 2015. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics 16:322.

69. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5(3):e9490.

70. Nofiani R, Philmus B, Nindita Y, Mahmud T. 2019. 3-Ketoacyl-ACP synthase (KAS) III homologues and their roles in natural product biosynthesis. MedChemComm 10(9):1501-1658.

71. Davies C, Heath RJ, White SW, Rock CO. 2000. The 1.8 Å crystal structure and active-site architecture of β-ketoacyl-acyl carrier protein synthase III (FabH) from Escherichia coli. Structure 8(2):185-195.

72. Röttig M, Rausch C, Kohlbacher O. 2010. Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families. PLoS Computational Biology 6(1):e1000636.

73. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. 2017. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics 33(20):3202-3210.

74. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. 2011. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Research 39:W362-W367.

75. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research 47:W81-W87.

76. Yang M, Fehl C, Lees K v., Lim E-K, Offen WA, Davies GJ, Bowles DJ, Davidson MG, Roberts SJ, Davis BG. 2018. Functional and informatics

analysis enables glycosyltransferase activity prediction. Nature Chemical Biology 14:1109-1117.

77. Pethe MA, Rubenstein AB, Khare SD. 2019. Data-driven supervised learning of a viral protease specificity landscape from deep sequencing and molecular simulations. Proceedings of the National Academy of Sciences 116(1):168-176.

78. Chen C-T, Yang E-W, Hsu H-J, Sun Y-K, Hsu W-L, Yang A-S. 2008. Protease substrate site predictors derived from machine learning on multilevel substrate phage display data. Bioinformatics 24(23):2691-2697.

79. Song J, Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC, Pike RN. 2012. PROSPER: An Integrated Feature-Based Tool for Predicting Protease Substrate Cleavage Sites. PLoS ONE 7(11):e50300.

80. Engström K, Nyhlén J, Sandström AG, Bäckvall J-E. 2010. Directed Evolution of an Enantioselective Lipase with Broad Substrate Scope for Hydrolysis of α-Substituted Esters. Journal of the American Chemical Society 132(20):7038-7042.

81. Cao D-S, Xiao N, Xu Q-S, Chen AF. 2015. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics 31(2):279-281.

82. Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. 2008. ChemmineR: a compound mining framework for R. Bioinformatics 24(15):1733-1740.

83. Atchley WR, Zhao J, Fernandes AD, Druke T. 2005. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences 102(8):6395-6400.

84. Osorio D, Rondón-Villarreal P, Torres R. 2015. Peptides: A Package for Data Mining of Antimicrobial Peptides. The R Journal 7(1):4-14.

85. Kuhn M. 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28(5):1-26.

86. Wright MN, Ziegler A. 2017. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77(1):1-17.

87. Wallace IM, Higgins DG. 2007. Supervised multivariate analysis of sequence groups to identify specificity determining residues. BMC Bioinformatics 8(135):1-12.

88. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10:845-858.

89. Binkowski TA. 2003. CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Research 31(13):3352-3355.

90. Liang G, Li Z. 2007. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. QSAR & Combinatorial Science 26(6):754-763.

91. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. 1985. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry 4(1):23-55.

92. Sukovich DJ, Seffernick JL, Richman JE, Hunt KA, Gralnick JA, Wackett LP. 2010. Structure, Function, and Insights into the Biosynthesis of a Head-to-Head Hydrocarbon in *Shewanella oneidensis* Strain MR-1. Applied and Environmental Microbiology 76(12):3842-3849.

93. Allemann MN, Shulse CN, Allen EE. 2019. Linkage of Marine Bacterial Polyunsaturated Fatty Acid and Long-Chain Hydrocarbon Biosynthesis. Frontiers in Microbiology 10:702.

94. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. 2002. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. Journal of Medicinal Chemistry 45(12):2615-2623.

95. Huffman MA, Fryszkowska A, Alvizo O, Borra-Garske M, Campos KR, Canada KA, Devine PN, Duan D, Forstater JH, Grosser ST, Halsey HM, Hughes GJ, Jo J, Joyce LA, Kolev JN, Liang J, Maloney KM, Mann BF, Marshall NM, McLaughlin M, Moore JC, Murphy GS, Nawrat CC, Nazor J, Novick S, Patel NR, Rodriguez-Granillo A, Robaire SA, Sherer EC, Truppo MD, Whittaker AM, Verma D, Xiao L, Xu Y, Yang H. 2019. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. Science 366(6470):1255-1259.

96. Meng G, Guo T, Ma T, Zhang J, Shen Y, Sharpless KB, Dong J. 2019. Modular click chemistry libraries for functional screens using a diazotizing reagent. Nature 574:86-89.

97. d'Espaux L, Mendez-Perez D, Li R, Keasling JD. 2015. Synthetic biology for microbial production of lipid-based biofuels. Current Opinion in Chemical Biology 29:58-65.

98. Wackett LP. 2011. Engineering microbes to produce biofuels. Current Opinion in Biotechnology 22(3):388-393.

99. Zargar A, Bailey CB, Haushalter RW, Eiben CB, Katz L, Keasling JD. 2017. Leveraging microbial biosynthetic pathways for the generation of 'drop-in' biofuels. Current Opinion in Biotechnology 45:156-163.

100. Lehmann J, Cheng T-Y, Aggarwal A, Park AS, Zeiler E, Raju RM, Akopian T, Kandror O, Sacchettini JC, Moody DB, Rubin EJ, Sieber SA. 2018. An Antibacterial β-Lactone Kills Mycobacterium tuberculosis by Disrupting Mycolic Acid Biosynthesis. Angewandte Chemie International Edition 57(1):348-353.

101. Wencewicz T. 2018. β-Lactam and β-Lactone Antibiotics from Plant Microbiomes. The FASEB Journal 32(51):257.1.

102. Böttcher T, Sieber SA. 2012. β-Lactams and β-lactones as activity-based probes in chemical biology. MedChemComm 3(4):408-417.

103. Bush K, Bradford PA. 2016. β-Lactams and β-Lactamase Inhibitors: An Overview. Cold Spring Harbor Perspectives in Medicine 6(8):a025247.

104. Tooke CL, Hinchliffe P, Bragginton EC, Colenso CK, Hirvonen VHA, Takebayashi Y, Spencer J. 2019. β-Lactamases and β-Lactamase Inhibitors in the 21st Century. Journal of Molecular Biology 431(18):3472-3500.

105. Yang Y, Li Y, Gu H, Dong M, Cai Z. 2020. Emerging agents and regimens for multiple myeloma. Journal of Hematology & Oncology 13:150.

106. Austin MB, Noel JP. 2003. The chalcone synthase superfamily of type III polyketide synthases. Natural Product Reports 20:79-110.

107. Fox AR, Soto G, Mozzicafreddo M, Garcia AN, Cuccioloni M, Angeletti M, Salerno JC, Ayub ND. 2014. Understanding the function of bacterial and eukaryotic thiolases II by integrating evolutionary and functional approaches. Gene 533(1):5-10.

108. Hong J, Park W, Seo H, Kim I-K, Kim K-J. 2020. Crystal structure of an acetyl-CoA acetyltransferase from PHB producing bacterium *Bacillus cereus* ATCC 14579. Biochemical and Biophysical Research Communications 533(3):442-448.

109. Robinson SL, Smith MD, Richman JE, Aukema KG, Wackett LP. 2020. Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. Synthetic Biology 5(1):ysaa4.

110. Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. Acta Crystallographica Section D Biological Crystallography 66(4):486-501.

111. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A,

Wilson KS. 2011. Overview of the CCP 4 suite and current developments. Acta Crystallographica Section D Biological Crystallography 67(4):235-242.

112. Gabrielson SW. 2018. SciFinder. Journal of the Medical Library Association 106.

113. SciFinder. SciFinder – Chemical Abstracts Service.

114. Kawaguchi A, Yoshimura T, Okuda S. 1981. A New Method for the Preparation of Acyl-CoA Thioesters. The Journal of Biochemistry 89(2):337-339.

115. Arora P, Vats A, Saxena P, Mohanty D, Gokhale RS. 2005. Promiscuous Fatty Acyl CoA Ligases Produce Acyl-CoA and Acyl-SNAC Precursors for Polyketide Biosynthesis. Journal of the American Chemical Society 127(26):9388-9389.

116. Peter D, Vögeli B, Cortina N, Erb T. 2016. A Chemo-Enzymatic Road Map to the Synthesis of CoA Esters. Molecules 21(4):517.

117. Bacher A, Stohler P, Weber W. 1996. Process for the production of lipstatin and tetrahydrolipstatin. EP-0803576-A2.

118. Tiwari S, Naveen Kumar CR, Sathyanathan D, Goel A, Iyer H. May 2008. Improved Fermentation Process for Higher Yield Coefficient of Lipase-Inhibitor with Respect to Consumed Fatty Acid. EP 2 268 822 B1.

119. Heitmann P. 1968. Reactivity of Sulfhydryl Groups in Micelles. A Model for Protein. European Journal of Biochemistry 5(3):305-315.

120. Krati N, Brembilla A, Lochon P. 1989. Unusual behaviour of a hydrophobic benzimidazolemethanethiol in micellar esterolysis. Tetrahedron Letters 30(1):2-14.

121. Tee OS, Yazbeck OJ. 2000. Transition state stabilization by micelles: thiolysis of *p* -nitrophenyl alkanoates in cetyltrimethylammonium bromide micelles. Canadian Journal of Chemistry 78(8):1100-1108.

122. Shinkai S, Kunitake T. 1976. Micellar Activation of Nucleophilic Reactivity of Coenzyme A and Glutathione toward *p* -Nitrophenyl Acetate. Bulletin of the Chemical Society of Japan 49(11):3219-3223.

123. Vlasova IM, Zhuravleva V v., Saletskii AM. 2013. Denaturation of bovine serum albumin under the action of cetyltrimethylammonium bromide, according to data from fluorescence analysis. Russian Journal of Physical Chemistry A 87:1027–1034.

124. Katre Y, Goyal N, Sharma R, Ajaya &, Singh K. 2013. Influence of cetyltrimethylammonium bromide/ sodium dodecyl sulfate micelles on the

oxidation of L-arginine by N-bromophthalimide in presence of HClO 4Indian Journal of Chemistry 52(6):732-738.

125. Purwono B, Dwi Wahyunngsih T. 2012. Micellar Catalytic Effect of Cetyltrimethylammonium Bromide on O-Allylation of Eugenol by Allyl BromideInternational Journal of Engineering & Technology IJET-IJENS 12(1):1-4.

126. Mir MA, Khan JM, Khan RH, Dar AA, Rather GM. 2012. Interaction of cetyltrimethylammonium bromide and its gemini homologue bis(cetyldimethylammonium)butane dibromide with xanthine oxidase. Journal of Physical Chemistry B 116(19):5711–5718.

127. Verma SK, Ghosh KK. 2011. Effect of cationic surfactants on the enzymatic activity of α-chymotrypsin. Kinetics and Catalysis 52(1):6-10.

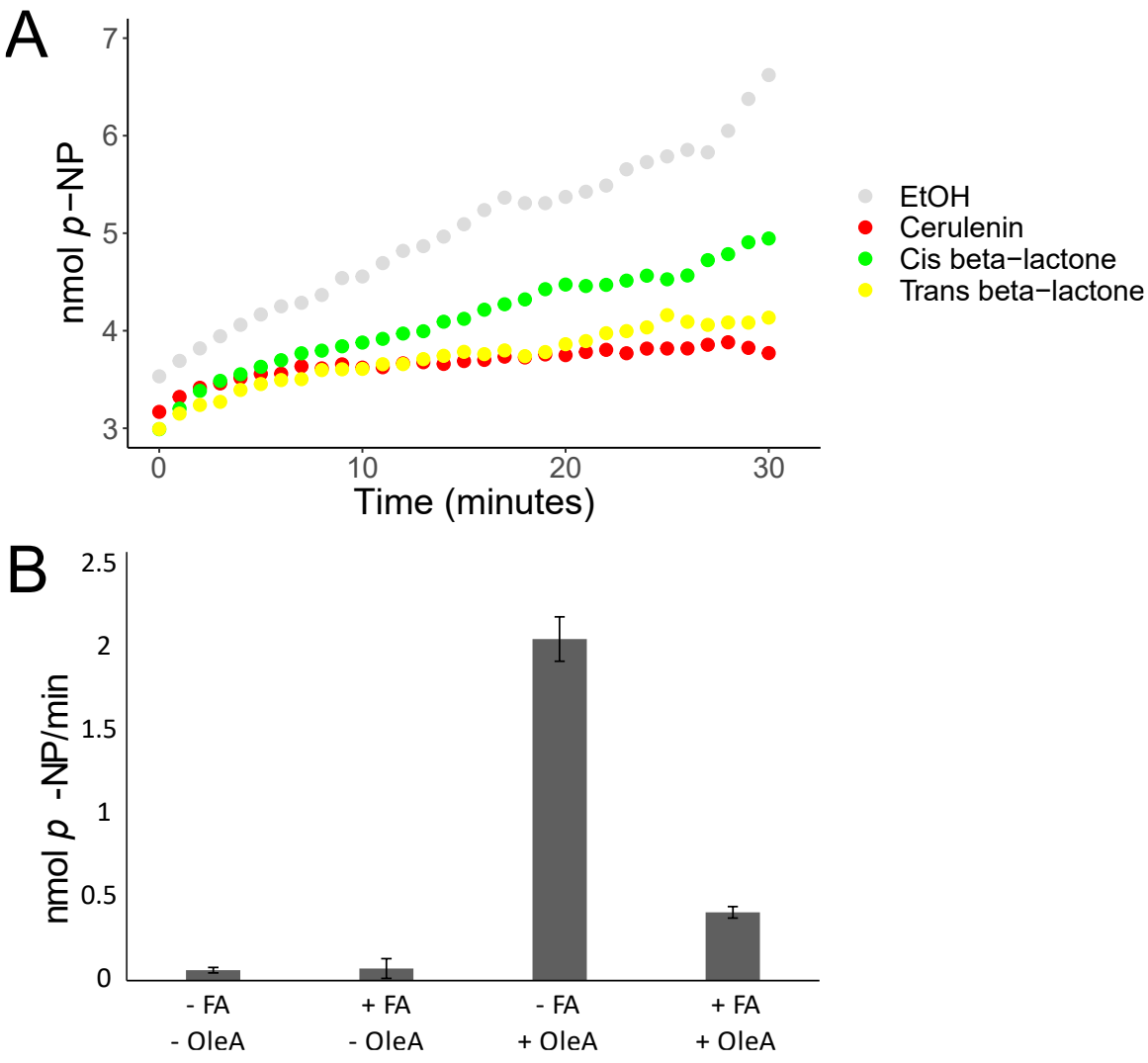
128. Puri NK, Crivelli E, Cardamone M, Fiddes R, Bertolini J, Ninham B, Brandon MR. 1992. Solubilization of growth hormone and other recombinant proteins from Escherichia coli inclusion bodies by using a cationic surfactant. Biochemical Journal 285(3):871-879.

129. Wang J, Lu D, Lin Y, Liu Z. 2005. How CTAB assists the refolding of native and recombinant lysozyme. Biochemical Engineering Journal 24(3):269–277.

130. Seddon AM, Curnow P, Booth PJ. 2004. Membrane proteins, lipids and detergents: not just a soap opera. Biochimica et Biophysica Acta (BBA) - Biomembranes 1666(1):105-117.

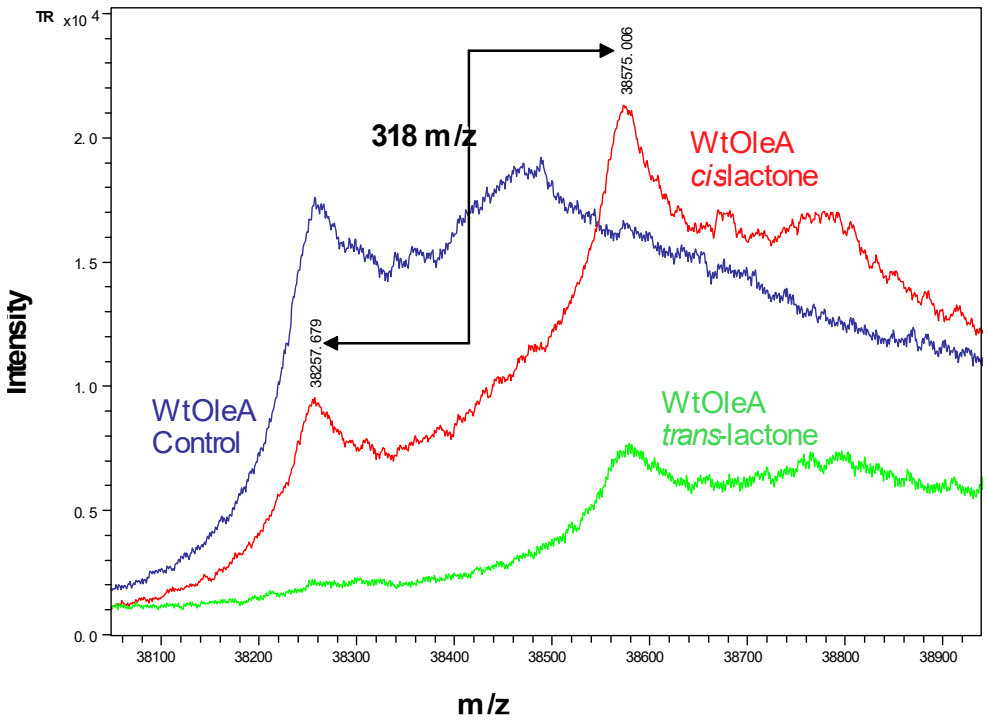
131. Umezawa H, Aoyagi T, Uotani K, Hamada M, Takeuchi T, Takahashi S. 1980. Ebelactone, an inhibitor of esterase, produced by actinomycetes. The Journal of Antibiotics 33(12):1594-1596.

Chapter 1 Supplemental

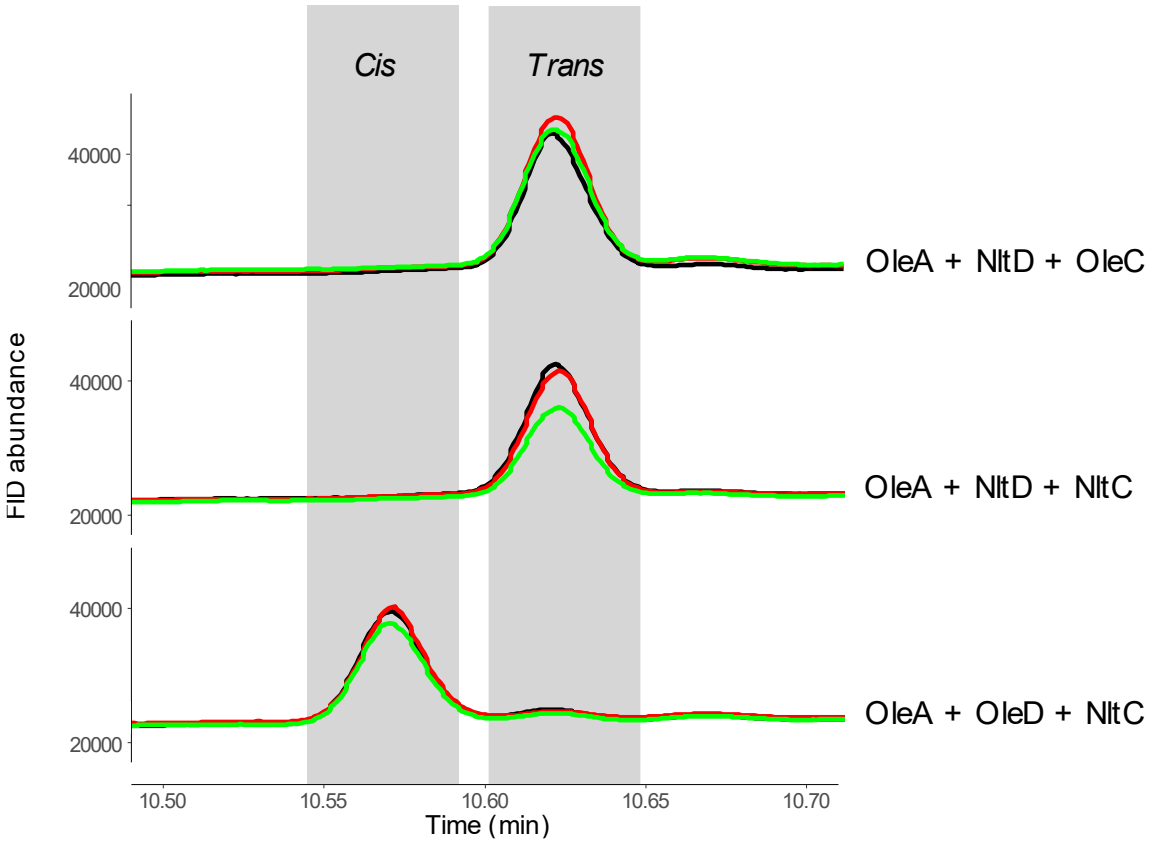


Supplemental Figure 1-S1. Inhibition of OleA by β-lactones. (A) Graph of WT OleA inhibited by addition of *cis* or *trans* β-lactone, as well as cerulenin, a known OleA inhibitor (32). 4 µg of OleA was incubated in 200 µL of 50 mM TrisHCl pH 8 + 5 % EtOH, with either 200 µM of cerulenin, *cis* β-lactone or *trans* β-lactone (R1, R2 = C9H19, C8H17). Solutions were pre-incubated at 37 C for 10 minutes, before addition of 200 µM of *p*-nitrophenyl laurate. Absorbance at was recorded at 410 nm, every minute for 30 minutes on a SpectraMax Plus384 microplate reader (Molecular Devices) at RT. All conditions were run in triplicate.

Conversion to nmol *p*-nitrophenol was calculated using extinction coefficient in Smith *et al.* (43) (B) β-lactone produced by OleADC + starting substrate inhibits OleA. 200 µL reactions in 50 mM TrisHCl pH 8 were run with 4 µg each of enzymes OleD, and OleC, along with an excess of $MgCl_2$, ATP and NADPH cofactors. Four conditions were then developed. A negative control (leftmost bar graph) was created with no other additions of myristoyl CoA (denoted in the graph as FA) or OleA. A second negative control (2^{nd} bar from left) was created with the addition of 100 µM of myristoyl CoA, but no addition of OleA. A positive control (3^{rd} bar from left) was created with the addition of 4 µg OleA, but no myristoyl CoA. Finally, the experimental condition (4^{th} bar from left) was created with the addition of 100 µM myristoyl-CoA and 4 µg of OleA. This completed the OleADC pathway and allowed for the creation of β-lactones that may inhibit OleA. All conditions were run in triplicate and allowed to react for 4 hours at RT. 100 µM *p*-nitrophenyl dodecanoate was then added to test for the activity of OleA following this reaction. Absorbance was measured at 410 in the colorimeter every minute for 30 minutes at 37 C. There was an 84.2% decrease in OleA activity between the positive control (no β-lactone present) and experimental condition (β-lactone formed by reaction of myristoyl-CoA with OleADC pathway).

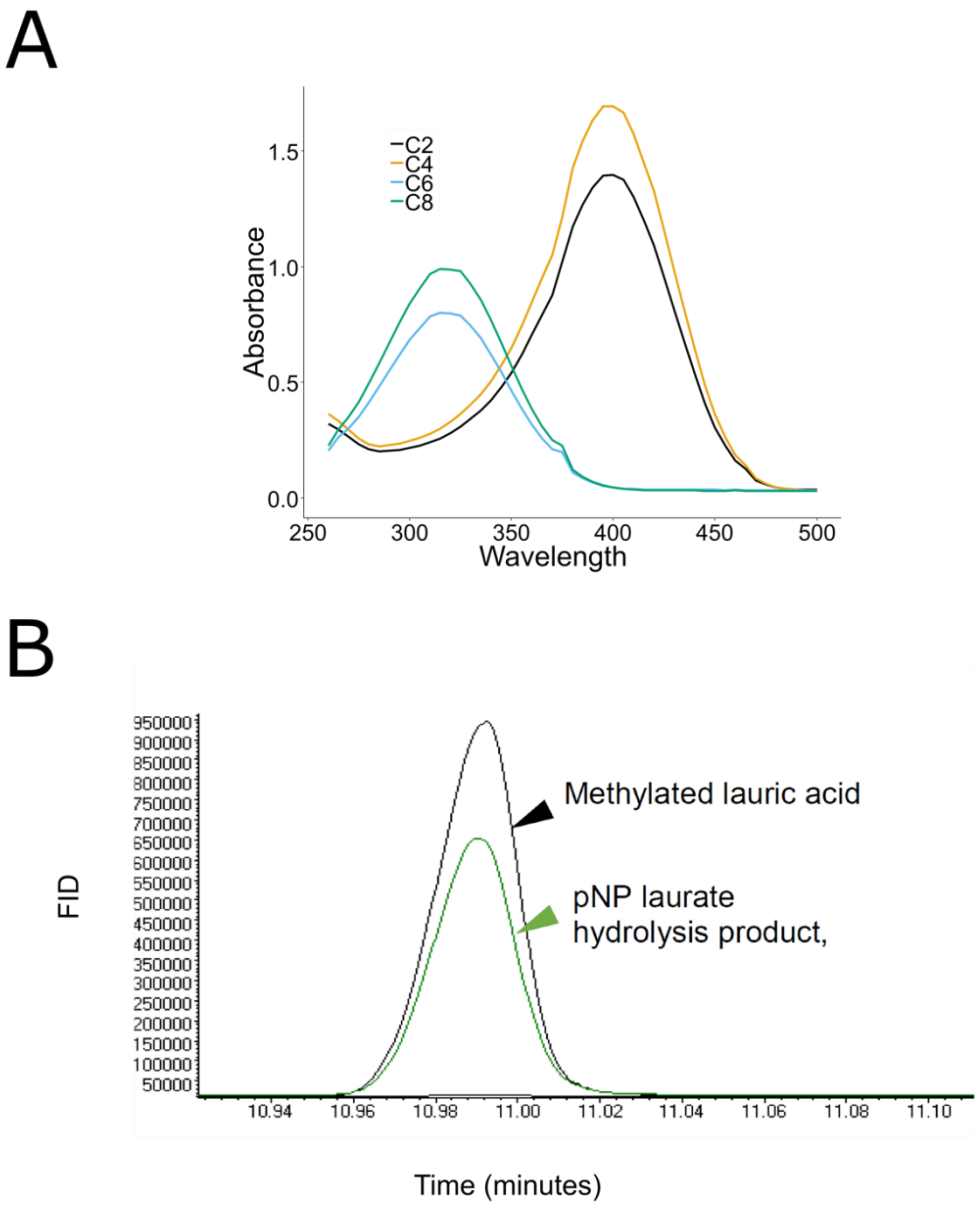


Supplemental 1-S2. MALDI-TOF of OleA inhibited by β-lactone. OleA was incubated in 500 µL of 50 mM TrisHCl with excess *cis* or *trans* β-lactone at RT overnight to allow for inhibition of OleA. This was then run through a MALDI-TOF procedure by the Center for Mass Spectronomy and Proteomics. Preprocessing was done using a Zip Tip protocol, adapted from Millipore sigma. C4 resin was used for protein desalting. A full protocol can be found on the website for the Center for Mass Spectronomy and Proteomics at the University of Minnesota (https://cbs.umn.edu/cmsp/protocols#ZipTip). The instrument used was the Autoflex speed MALDI TOF/TOF System (Bruker). A shift in peak was seen corresponding to 318 m/z, which represents the bound β-lactone. This experiment was done in collaboration with Dr. Serina Robinson and the Center for Mass Spectronomy and Proteomics.

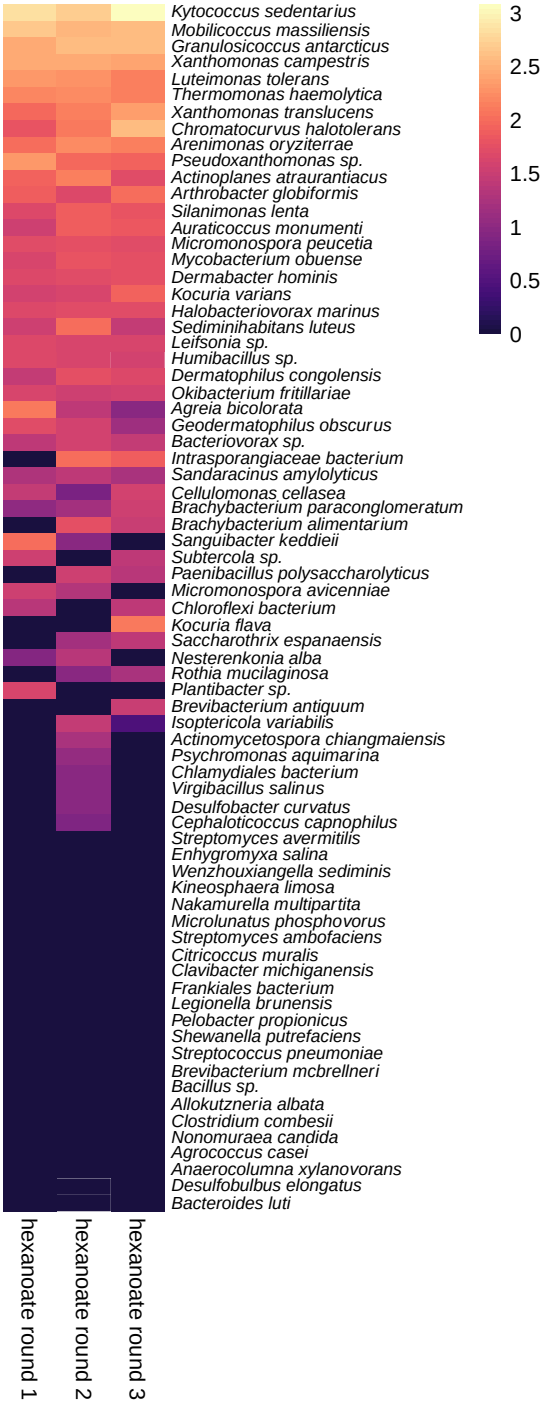


Supplemental Figure 1-S3. Manipulation of β-lactone stereochemistry. 500 µL reactions in 50 mM TrisHCl pH 8 were run in triplicate containing 9 µg of OleA + 9 µg of OleD or NltD + 9 µg of OleC or NltC with an excess of $MgCl_2$, ATP, and NADPH as cofactors. 200 µM of Decanoyl CoA was added to each condition and allowed to react at RT overnight. The resulting β-lactone was extracted in 500 µL MTBE, and run on GC-MS following the protocol detailed in Smith *et al* (43). The β-lactone product (R1, R2 = C9H19, C8H17) under the high heat of the GC decarboxylate readily into olefin, which is seen above *Cis*-olefin is detected at 10.55 – 10.60 minutes, and *trans* olefin is detected at 10.60 – 10.65 minutes. OleA + OleD + OleC is a known reaction published by Christenson *et al* (6), and so was not tested in this experiment.

Chapter 2 Supplemental



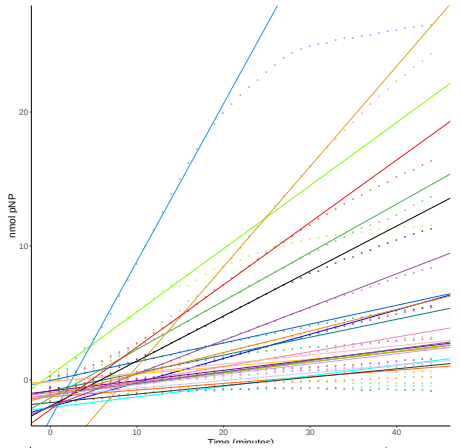
Supplemental Figure 2-S1. Demonstration of hydrolysis of *p*-nitrophenyl laurate. (A) Release of *p*-nitrophenol demonstrated by UV/vis spectroscopy; black curve of OleA enzyme product, gold curve is standard *p*-nitrophenol; blue curve is the OleA enzyme product after addition of HCl and green is the *p*-nitrophenol standard after HCl addition. (B) Gas chromatograph of OleA reaction mixture or standard lauric acid after methylation by diazomethane.



Supplemental Figure 2-S2. Heatmap of relative enzyme activity ($\log_{10}$ of nmol pNP produced over the course of one hour by a *E. coli* BL21 culture with an OD of 1.0) of 73 OleA enzymes across three different replicates. Enzyme activity ranges from high activity (yellow) to no activity (dark purple). Results reveal general

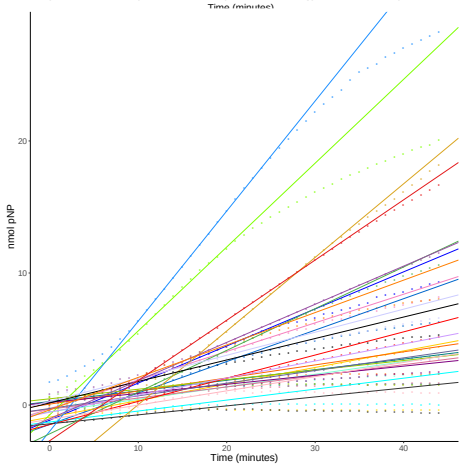
reproducibility of this whole-cell method for enzymes with high activity and lower reproducibility for enzymes with weak activity.

Biological replicate 1



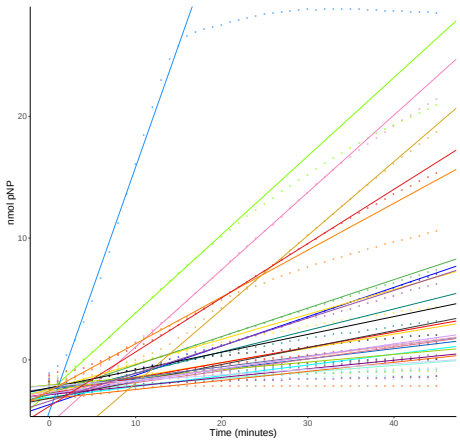
Source organism	Enzyme activity
Kytococcus sedentarius	2.85
Mobilicoccus massiliensis	2.65
Granulosicoccus antarcticus	2.45
Xanthomonas campestris	2.45
Luteimonas tolerans	2.34
Pseudoxanthomonas	2.3
Thermomonas haemolytica	2.17
Arenimonas oryziterrae	2.04
Xanthomonas translucens	1.99
Actinoplanes atraurantiacus	1.93
Arthrobacter globiformis	1.9
Chromatocurvus halotolerans	1.82
Micromonospora peucetia	1.71
Humibacillus sp. DSM 29435	1.69
Halobacteriovorax marinus	1.68
Leifsonia sp. Leaf325	1.67
Silanimonas lenta	1.66
Dermabacter hominis	1.66
Mycobacterium obuense	1.61
Kocuria varians	1.58
Sediminihabitans luteus	1.54
Auraticoccus monumenti	1.52
Dermatophilus congolensis	1.47

Biological replicate 2



Source organism	Enzyme activity
Kytococcus sedentarius	2.7
Granulosicoccus antarcticus	2.58
Mobilicoccus massiliensis	2.52
Xanthomonas campestris	2.44
Luteimonas tolerans	2.28
Thermomonas haemolytica	2.24
Arenimonas oryziterrae	2.23
Xanthomonas translucens	2.17
Actinoplanes atraurantiacus	2.15
Chromatocurvus halotolerans	2.11
Intrasporangiaceae bacterium URHB0013	2.02
Sediminihabitans luteus	2.02
Pseudoxanthomonas	1.99
Silanimonas lenta	1.89
Auraticoccus monumenti	1.88
Mycobacterium obuense	1.81
Dermatophilus congolensis	1.77
Micromonospora peucetia	1.76
Brachybacterium alimentarium	1.76
Dermabacter hominis	1.7
Halobacteriovorax marinus	1.7
Arthrobacter globiformis	1.68
Humibacillus sp. DSM 29435	1.64
Leifsonia sp. Leaf325	1.62
Kocuria varians	1.61

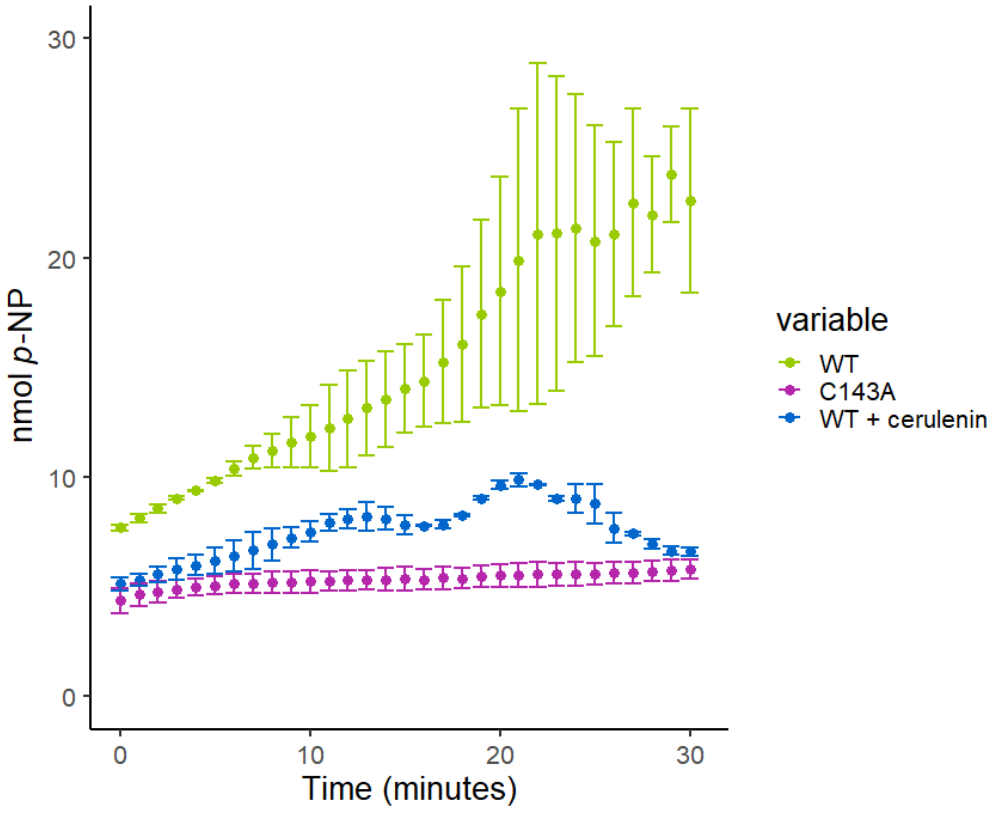
Biological replicate 3



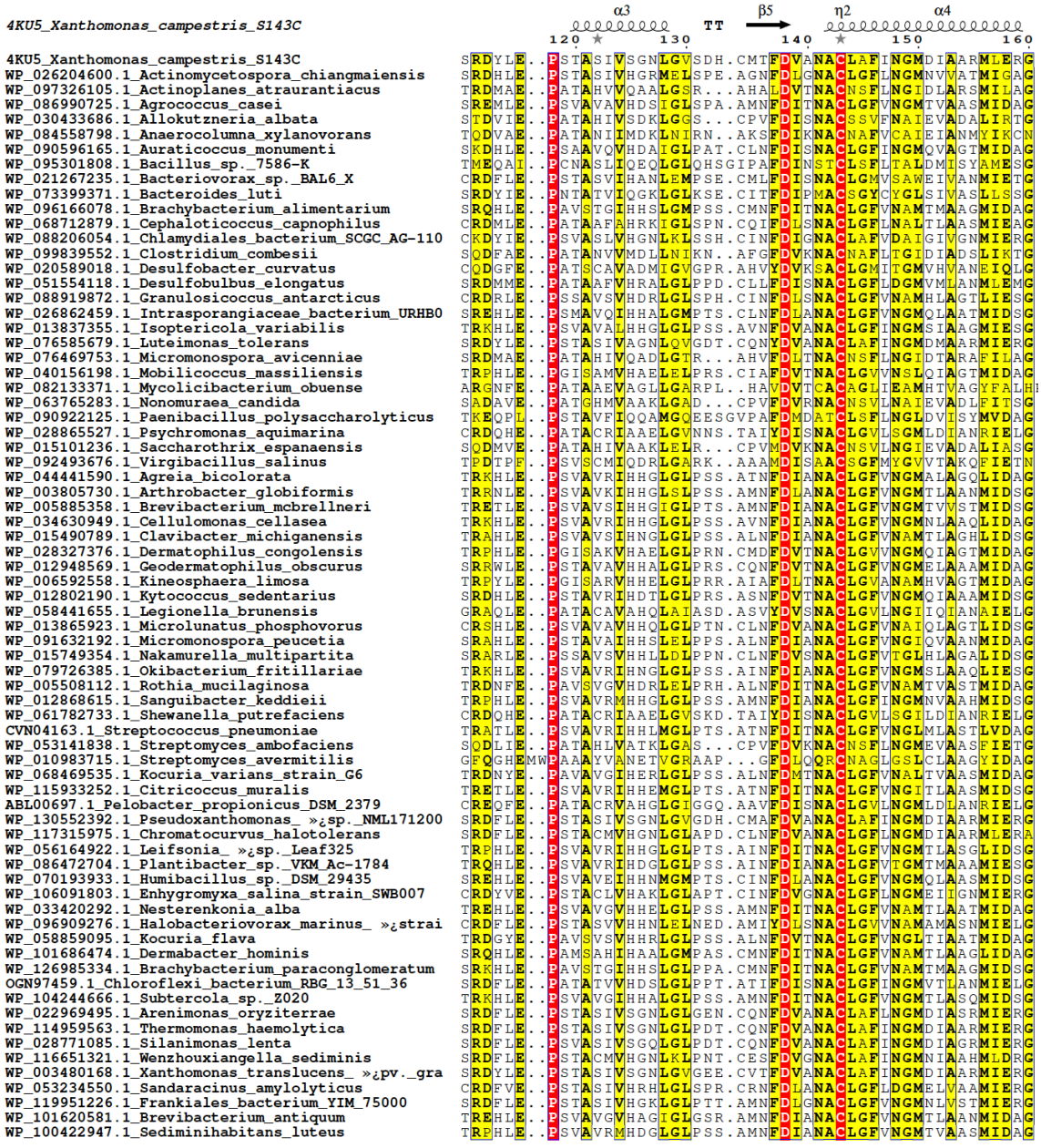
Source organism	Enzyme activity
Kytococcus sedentarius	3.08
Granulosicoccus antarcticus	2.59
Chromatocurvus halotolerans	2.58
Mobilicoccus massiliensis	2.56
Xanthomonas campestris	2.43
Xanthomonas translucens	2.36
Arenimonas oryziterrae	2.16
Luteimonas tolerans	2.15
Thermomonas haemolytica	2.13
Kocuria flava	2.09
Arthrobacter globiformis	2.02
Pseudoxanthomonas	1.94
Kocuria varians	1.92
Intrasporangiaceae bacterium URHB0013	1.88
Auraticoccus monumenti	1.85
Silanimonas lenta	1.79
Dermabacter hominis	1.77
Mycobacterium obuense	1.76
Halobacteriovorax marinus	1.72
Micromonospora peucetia	1.72
Actinoplanes atraurantiacus	1.7
Dermatophilus congolensis	1.65
Leifsonia sp. Leaf325	1.63
Humibacillus sp. DSM 29435	1.58
Brachybacterium alimentarium	1.51
Sediminihabitans luteus	1.44

Supplemental Figure 2-S3. Graph of active OleA enzymes in whole cell *p*-NP assays across all 3 replicates, along with a reference line for calculated slope.

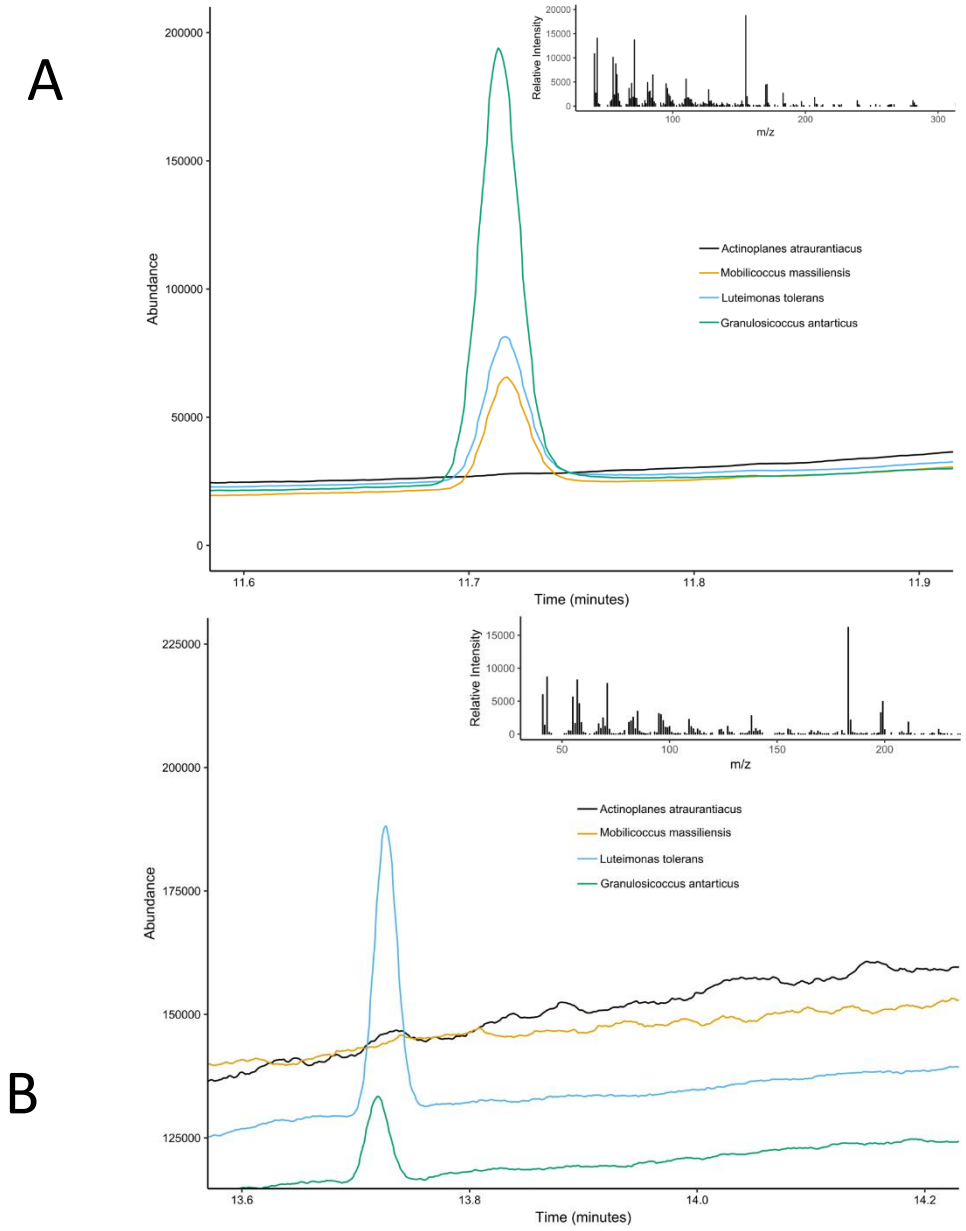
Active organisms for each replicate are listed besides the table as well as rate of enzyme activity in nmol p-NP / OD 1.0 / hour.



Supplemental Figure 2-S4. 4 µg of OleA WT (green) and OleA C143A (magenta) and WT with 25 µM cerulenin (blue) were assayed with 200 µM *p*-nitrophenyl dodecanoate in 50 mM Tris HCl pH 8 at 37 C. Cerulenin was added prior to the assay and the plate was allowed to incubate for 10 minutes at 37 C.

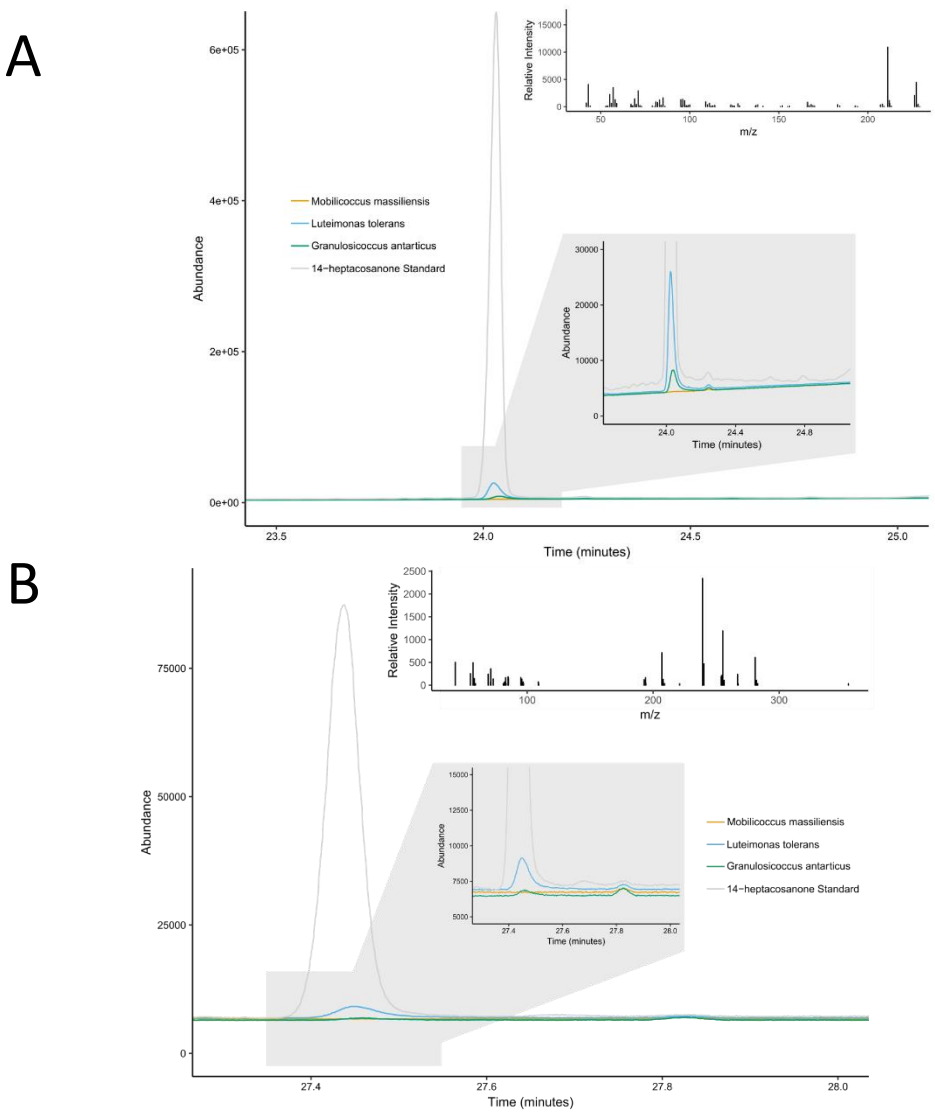


Supplemental Figure 2-S5. Structure-based multiple sequence alignment of OleA proteins screened in this study. The numbering and structural elements depicted are based on the X-ray structure of the *X. campestris* OleA. Glutamate 117 is highly conserved, proline 118 and cysteine 143 are completely conserved.



Supplemental Figure 2-S6. A.) GC-FID/MS split chromatogram of an extract from reaction of OleA proteins from different sources with decanoyl-CoA showing 10-nonadecanone, the stable decarboxylated product of the Claisen condensation that can be detected by GC. Black is *A. atraurantiacus*, orange is *M. massiliensis,* blue is *L. tolerans,* green is *G. antarticus*. Inset shows the MS spectrum of the peak (from the representative *Luteimonas tolerans* OleA). All mass spectra from reactions showing a peak gave the characteristic fragments for 10-noadecanone. **B.)** GC-FID/MS split chromatogram of an extract from

reaction of OleA proteins from different sources with lauryl-CoA showing 12-tricosanone, the stable decarboxylated product of the Claisen condensation that can be detected by GC. Black is *A. atraurantiacus*, orange is *M. massiliensis,* blue is *L. tolerans,* green is *G. antarticus*. Inset shows the MS spectrum of the eluted standard peak. All mass spectra from reactions showing a peak gave the characteristic fragments for 12-tricosanone.



Supplemental Figure 2-S8. A.) GC-FID/MS split chromatogram of an extract from reaction of OleA proteins from different sources with myristoyl-CoA showing 14-heptacosanone, the stable decarboxylated product of the Claisen condensation that can be detected by GC. Orange is *M. massiliensis,* blue is *L. toleran,* green is *G. antarticus,* grey is the chemical standard. *A. atraurantiacus* was not included due to inactivity and insolubility. Inset shows the MS spectrum of the eluted standard peak. All mass spectra from reactions showing a peak gave the characteristic fragments for 14-heptacosanone. **B.)** GC-FID/MS split chromatogram of an extract from reaction of OleA proteins from different sources with palmitoyl-CoA showing 16-hentriacontanone, the stable decarboxylated

product of the Claisen condensation that can be detected by GC. Orange is *M. massiliensis,* blue is *L. tolerans,* green is *G. antarcticus,* grey is the chemical standard. *A. atraurantiacus* was not included due to inactivity and insolubility. Inset shows the MS spectrum of the eluted standard peak. All mass spectra from reactions showing a peak gave the characteristic fragments for 16-hentriacontanone.

Accession Number	Organism Name	%ID to X. campestris OleA
WP_003480168	*Xanthomonas_translucens*	87.9
WP_130552392	*Pseudoxanthomonas_sp.*	87.3
WP_022969495	*Arenimonas_oryziterrae*	80.8
WP_114959563	*Thermomonas_haemolytica*	79.9
WP_028771085	*Silanimonas_lenta*	77.8
WP_076585679	*Luteimonas_tolerans*	77.2
WP_117315975	*Chromatocurvus_halotolerans*	59.8
WP_116651321	*Wenzhouxiangella_sediminis*	58.6
WP_053234550	*Sandaracinus_amylolyticus*	53.0
WP_106091803	*Enhygromyxa_salina*	52.1
WP_119951226	*Frankiales_bacterium*	49.1
WP_088206054	*Chlamydiales_bacterium*	46.2
WP_013837355	*Isosphaericola_variabilis*	45.5
WP_101620581	*Brevibacterium_antiquum*	45.4
WP_079726385	*Okibacterium_fritillariae*	45.2
WP_012802190	*Kytococcus_sedentarius*	45.2
WP_026204600	*Actinomycetospora_chiangmaiensis*	44.8
WP_100422947	*Sediminihabitans_luteus*	44.4
WP_026862459	*Intrasporangiaceae_bacterium*	44.3
WP_015490789	*Clavibacter_michiganensis*	43.9
WP_070193933	*Humibacillus_sp.*	43.7
WP_012868615	*Sanguibacter_keddieii*	43.6
WP_044441590	*Agreia_bicolorata*	43.3
WP_086990725	*Agrococcus_casei*	43.0
WP_090596165	*Auraticoccus_monumenti*	42.9
WP_013865923	*Microlunatus_phosphovorus*	42.9
WP_012948569	*Geodermatophilus_obscurus*	42.9
WP_091632192	*Micromonospora_peucetia*	42.6
WP_034630949	*Cellulomonas_cellasea*	42.4
WP_056164922	*Leifsonia_sp.*	42.4
WP_088919872	*Granulosicoccus_antarcticus*	42.3
WP_003805730	*Arthrobacter_globiformis*	42.2
WP_086472704	*Plantibacter_sp.*	42.0
WP_033420292	*Nesterenkonia_alba*	41.1
WP_005885358	*Brevibacterium_mcbrellneri*	40.9
WP_104244666	*Subtercola_sp.*	40.6

Accession Number	Organism Name	%ID to X. campestris OleA
WP_101686474	*Dermabacter_hominis*	40.5
WP_058859095	*Kocuria_flava*	39.6
WP_126985334	*Brachybacterium_paraconglomeratum*	39.6
WP_068469535	*Kocuria_varians*	39.4
WP_015749354	*Nakamurella_multipartita*	39.3
CVN04163	*Streptococcus_pneumoniae*	38.8
WP_115983252	*Citricoccus_muralis*	38.6
WP_096909276	*Halobacteriovorax_marinus*	38.2
WP_068712879	*Cephaloticoccus_capnophilus*	38.1
WP_096166078	*Brachybacterium_alimentarium*	38.0
WP_028327376	*Dermatophilus_congolensis*	37.8
WP_040156198	*Mobilicoccus_massiliensis*	37.4
OGN97459	*Chloroflexi_bacterium*	37.0
WP_005508112	*Rothia_mucilaginosa*	36.9
WP_006592558	*Kineosphaera_limosa*	36.8
WP_021267235	*Bacteriovorax_sp.*	36.6
WP_051554118	*Desulfobulbus_elongatus*	36.6
ABL00697	*Pelobacter_propionicus*	36.1
WP_028865527	*Psychromonas_aquimarina*	35.0
WP_061782733	*Shewanella_putrefaciens*	34.1
WP_058441655	*Legionella_brunensis*	32.6
WP_020589018	*Desulfobacter_curvatus*	32.5
WP_030433686	*Allokutzneria_albata*	29.3
WP_097326105	*Actinoplanes_atraurantiacus*	29.1
WP_053141838	*Streptomyces_ambofaciens*	29.0
WP_076469753	*Micromonospora_avicenniae*	28.3
WP_010983715	*Streptomyces_avermitilis*	27.1
WP_099839552	*Clostridium_combesii*	27.0
WP_015101236	*Saccharothrix_espanaensis*	26.5
WP_082133371	*Mycobacterium_obuense*	25.9
WP_092493676	*Virgibacillus_salinus*	25.4
WP_090922125	*Paenibacillus_polysaccharolyticus*	25.1
WP_063765283	*Nonomuraea_candida*	24.9
WP_084558798	*Anaerocolumna_xylanovorans*	24.8
WP_073399871	*Bacteroides_luti*	24.8
WP_095301808	*Bacillus_sp.*	24.3

Supplemental Table 2-S1. Accession number, organism name, and percent amino acid sequence to *X. campestris* OleA of all OleAs expressed and tested in the present assay. The percent amino acid sequence identity was calculated using BLAST.

Supplemental Table 2-S2. Comparison of *in vitro* rates of *p*-nitrophenyl ester hydrolysis by OleA and other hydro enzymes

Enzyme	Source	(Pfam ID)	Protein family	Reaction	Active Site Nucleophile	Relative Rate (%)		
						C6	C10	C14
OleA (β-keto acid synthase)	*Xanthomonas campestris*	(PF08545)	3-Oxoacyl-ACP synthase III	Transferase (Synthase)	Cysteine	100	100	100
Intestinal Lipase	*Sus domesticus*	(PF00151)	Triglyceride Lipase	Ester Hydrolysis	Serine	13	89	8
Papain	*Ananas comosus*	(PF00112)	Peptidase C1	Amide Hydrolysis	Cysteine	<1	<1	<1
Chymotrypsin-α	*Sus domesticus*	(PF00089)	Trypsin	Amide Hydrolysis	Serine	100	8	4
Arylaceto-nitrilase	*Alcaligenes faecalis*	(PF00795)	CN Hydrolase	Nitrile Hydration	Cysteine	2	<1	7

Source organism	Enzyme activity
Kytococcus sedentarius	2.88
Mobilicoccus massiliensis	2.58
Granulosicoccus antarcticus	2.54
Xanthomonas campestris	2.44
Luteimonas tolerans	2.25
Thermomonas haemolytica	2.18
Xanthomonas translucens	2.17
Chromatocurvus halotolerans	2.17
Arenimonas oryziterrae	2.14
Kocuria flava	2.09
Pseudoxanthomonas	2.08
Intrasporangiaceae bacterium URHB0013	1.95
Actinoplanes atraurantiacus	1.93
Arthrobacter globiformis	1.87
Silanimonas lenta	1.78
Auraticoccus monumenti	1.75
Micromonospora peucetia	1.73
Mycobacterium obuense	1.73
Dermabacter hominis	1.71
Kocuria varians	1.7
Halobacteriovorax marinus	1.7
Sediminihabitans luteus	1.67
Leifsonia sp. Leaf325	1.64
Humibacillus sp. DSM 29435	1.63
Dermatophilus congolensis	1.63
Brachybacterium alimentarium	1.63

Supplemental Table 2-S3. Average enzyme activity ($\log_{10}$ of nmol pNP produced over the course of one hour by a *E. coli* BL21 culture with an OD of 1.0) for 73 OleAs.

Chapter 3 Supplemental

Supplemental Methods

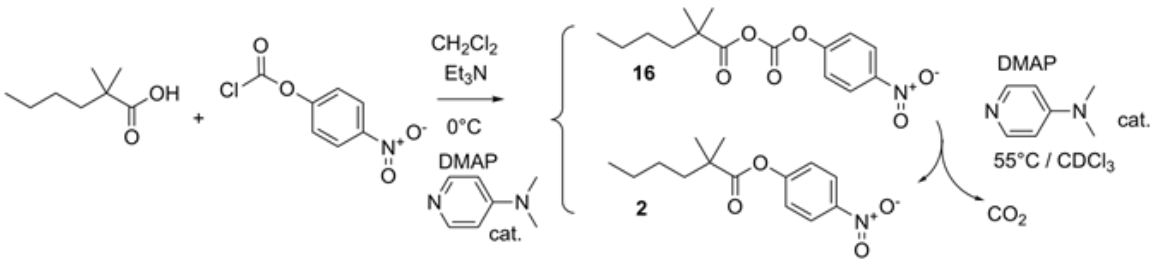
General: 1H NMR spectra were acquired at 400 MHz on a Varian Inova 400 spectrometer. Gas chromatography/mass spectrometry data were acquired on Agilent Tech 7890a GC System with a 30 m x 0.25 mm i.d. x 0.25 μm DB1-MS capillary column with outflow split to flame ionization and MS detectors. The MS (Agilent Tech 5975c) normally operated at 70ev, but also was operated with overloaded injections (up to 1% MTBE solutions) and at 10ev to detect very weak molecular ions. **Molecular ions, m/z, are denoted with boldface.** *Acylium ions, RCO+ are denoted with italics.* Retention times (temperatures) are not absolute but are significant in relation to other retention times. For more polar, higher-boiling point 4-nitrophenyl esters, direct injection electrospray ionization MS (instrument) in methanol succeeded at detecting molecular ions as M + H+ or + Na+ cations.

4-Nitrophenyl Esters of Carboxylic Acids. The general procedure described by Engström et al. (1) was used on ≈ 0.6 mmol scale. *Procedure A* is an example of this general procedure, which is applied to the esterification of normal-chain carboxylic acids (unbranched at the α- carbon). *Procedure B* is a modified procedure applicable for the esterification of 2,2-dimethylhexanoic acid, that describes how mixed anhydrides of 4-nitrophenyl carbonic acid and carboxylic acids, which are often minor products in *procedure A,* can be converted to 4-nitrophenyl esters of carboxylic acids.

4-Nitrophenyl 6-Azidohexanoate (1, *Procedure A*). To a magnetically stirred dry 10mL 14/20 RB flask fitted with a 14/20 inlet and additional sidearm inlet to allow nitrogen flush of the reactor, was charged 89 mg (0.57 mmol) of 6-azidohexanoic acid (Cayman Chem. Co., >98%), 1.5 mL of dry dichloromethane (DCM, mixture cooled to 0°C) and 114 mg (0.57 mmol) of 4-nitrophenyl chloroformate. After 10 min, a solution of 7 mg of 4-dimethylaminopyridine

(DMAP, ~10 mol%) and 63 mg (slight excess) of triethylamine in 1.5 mL DCM was added over 1 minute. This solution was stirred at 0°C for 2 h, then DCM was added to fill the reactor. The reaction mixture was stirred with 1 drop of 6N aq. HCl and the pH of the aqueous phase was raised to 6 with 1N aq. NaOH. The aqueous phase was discarded and the organic phase was concentrated, then washed (DCM) through a column of 2 g of silica gel/DCM. The UV active eluate was collected in 2 fractions, 20 mL each, which concentrated to 104 and 45 mg of thin colorless oils. 1H NMR and GC/MS analyses indicate that the first concentrated fraction of the UV-active eluate (liquid) which was used in the enzyme activity screens, contains ~0.5 % DCM and is ≥98% pure **1**. The second fraction is only slightly less pure. Yield ≈ 94%. GC/MS: 14.86 min (249°C, 70ev); overloaded and with 10ev ms: **m/z 278 (0.04%)**, *acylium (C6H10N3O+) 140 (50%)*, base fragment 69 (C2H3N3+ or C5H9+ or C4H5O+, 100%). 1H NMR (≥98%): 8.28 (m, 2H), 7.28 (m, 2H), 3.32 (t, 2H), 2.63 (t, 2H), 1.80 (m, 2H), 1.68 (m, 2H) and 1.56 – 1.48 ppm (m, 2H + H2O).

4-Nitrophenyl 2,2-dimethylhexanoate (2, *Procedure A*). Mixed anhydrides of 4-nitrophenyl carbonic acid and carboxylic acids are sometimes minor products when *procedure A* is being used to esterify linear carboxylic acids (acids that are unbranched at the α-carbon to the carboxyl carbon). *Procedure A* applied to the esterification of 2,2-dimethylhexanoic acid produces a considerable amount the mixed anhydride of 4-nitrophenyl carbonic acid (**16**) and 2,2- dimethylhexanoic acid in addition to the normally expected **2** (Scheme 1).



Scheme 3-1. Synthesis of **2** and **16**

The starting acid, 2,2-dimethylhexanoic acid (TCI, >98%) analyzed by 1H NMR is consistent with ~98% purity. Charges were 84.4 mg (0.585 mmol) of the acid, 2.0 mL of dry DCM (mixture cooled to 0°C), and 120 mg (0.595 mmol) of 4-nitrophenyl chloroformate. After 5 min, a solution of 6.8 mg of DMAP and 90 µL (66 mg, excess) triethylamine in 1.5 mL dry DCM was added. After 2 h at 0°C, the yellow reaction mixture was worked up as described above. Two fractions of UV active DCM eluates off the column gave 90 mg of thick colorless oil and 48 mg of crystalline solids. 1H NMR on a sample of the second eluate solids show the closely related **2** and **16** in 1.0: 0.6 mole ratio. For the first eluate this ratio is 1.0 : 0.26. Refluxing a concentrated solution (~1/2 mL of DCM) of 3.1 mg of DMAP and 42 mg of the second eluate solids and concentrating (N2) the warm liquid residue converted nearly all (2% remains) of **16** to **2**. DMAP was removed by filtering this DCM solution plus DCM washes through 0.4 g of silica gel to recover ~90% yield of >98% pure liquid **2**. 1H NMR of **16** is similar to that of **2**: [**16** vs **2**] 8.35 vs 8.27 (m, 2H), 7.505 vs 7.24 (m, 2H), 1.55 vs 1.685 (m, 2H), 1.4-1.24 vs 1.38-1.28 (m, 10H) and 0.905 vs 0.935 ppm (t, 3H). GC/MS for **2**: 11.74 min (217 °C), MS detector operating at 70 ev: **no molecular ion**, *acylium (C8H15O+) 127 (10%)*, C7H15+ 99 (70%) and base fragment m/z 57(C4H9+, 100%). Reacquired with MS at 10 ev, with overloaded GC shows only a very weak **molecular ion at m/z 265**.

4-Nitrophenyl 2,2-dimethylhexanoate (2, *Procedure B*). Modifying *procedure A* by omitting DMAP in this reaction, the mixed anhydride **16** is produced in high yield. As above, clean decarboxylation of **16** is catalyzed by DMAP. To the reactor described in *Procedure A* was charged 82 mg of 2,2-dimethylhexanoic acid, 2.0 mL of dry DCM, 117.8 mg of 4-nitrophenyl chloroformate (mixture cooled to 0°C) and 90 µL of triethylamine. The reaction mixture stirred 30 min at 0°C, then was stored under N2 at -18°C overnight. Work-up as described in *procedure A* produced three DCM fractions (10, 10 and 5 mL) which gave residues of 76, 82 and 6 mg respectively. All three residues are crystalline solids, 1 and 2 melt 128.5-133°C and 128-133°C. 1H NMR shows that fraction 3 is much

less pure and that fraction 1 is 98% pure for **16** (and 2% of **2**), *cf.* above: 8.355 (m, 2H), 7.51 (m, 2H), 1.5 (t, 2H), 1.30-1.28 (m, 4H), 1.23 (s, 6H) and 0.89 ppm (t, 3H). Decarboxylation of this product is described above (*Procedure A)*. 1H NMR, GC/MS and m.p. data on other 4-nitrophenyl esters of carboxylic acids not commercially-available or previously reported in the open literature are described below:

4-Nitrophenyl 2-(2-butoxyethoxy)acetate (3). Liquid at room temperature. MS (direct injection electrospray ionization: m/z 320 (M + Na+), little or no m/z 298 (M + H+). 1H NMR (~97% pure, ~3 mol % 4-nitrophenol): 8.29 (m, 2H), 7.33 (m, 2H), 4.47 (s, 2H), 3.83 (m, 2H), 3.67 (m, 2H), 3.49 (t, 2H), 1.58 (m, 2H + H2O), 1.37 (m, 2H) and 0.92 ppm (t, 3H).

4-Nitrophenyl 7-phenylheptanoate (4). m.p. 56.0 – 57.0°C. GC/MS: 18.43 min (284°C), overloaded, 10ev: **m/z 327 (0.2%)**, *acylium (C13H17O+) 189 (40%)*, base 91 (C7H7+, 100%). 1H NMR (≥98%): 8.27 (m, 2H), 7.265 (m, 2H), 7.30-7.24 (m, 2H), 7.20-7.16 (m, 3H), 2.64-2.57 (m, 4H overlapping t), 1.76 (m, 2H), 1.66 (m, 2H) and 1.5-1.36 ppm (m, 4H).

4-Nitrophenyl 6-heptynoate (5). m.p. 54.5 – 55.5°C. GC/MS: 12.31 min (223°C), overloaded, 10ev: **m/z 247 (0.05%)**, *acylium (C7H9O+) 109 (75%)*, base 81 (C6H9+, 100%). 1H NMR (≥98%): 8.27 (m, 2H), 7.2x (m, 2H), 2.65 (t, 2H), 2.28 (dt, J=2.6, 7.0 Hz, 2H), 1.99 (7, J=2.6 Hz, 1H), 1.90 (m, 2H) and 1.67 ppm (m, 2H).

4-Nitrophenyl 3-cyclopentylpropanoate (6). m.p. 50.0 – 51.0°C. GC/MS: 13.88 min (239°C),overloaded, 10ev: **m/z 263 (0.3%),** *acylium (C8H13O+) 125 (100%)*. 1H NMR (≥98%): 8.27 (m, 2H), 7.28 (m, 2H), 2.62 (distorted t, 2H), 1.9-1.75 (m, 5H), 1.7-1.5 (m, 4H) and 1.2-1.1 ppm (m, 2H).

4-Nitrophenyl 3-(4-chlorophenoxy)propanoate (7). m.p. 98 – 99°C. GC/MS: 17.70 min (277°C) – mostly decomposed, small peak survives GC – overloaded, 10ev, **m/z 321 and 323 (20%)**, *acylium (C9H9ClO2+) 183 and 185 (80%)*, base (C7H6ClO+, 100%) 141 and 143. 1H NMR (≥95%): 8.28 (m, 2H), 7.35-7.22 (m, 4H), 6.87 (m, 2H), 4.34 (t, 2H) and 3.09 ppm (t, 2H).

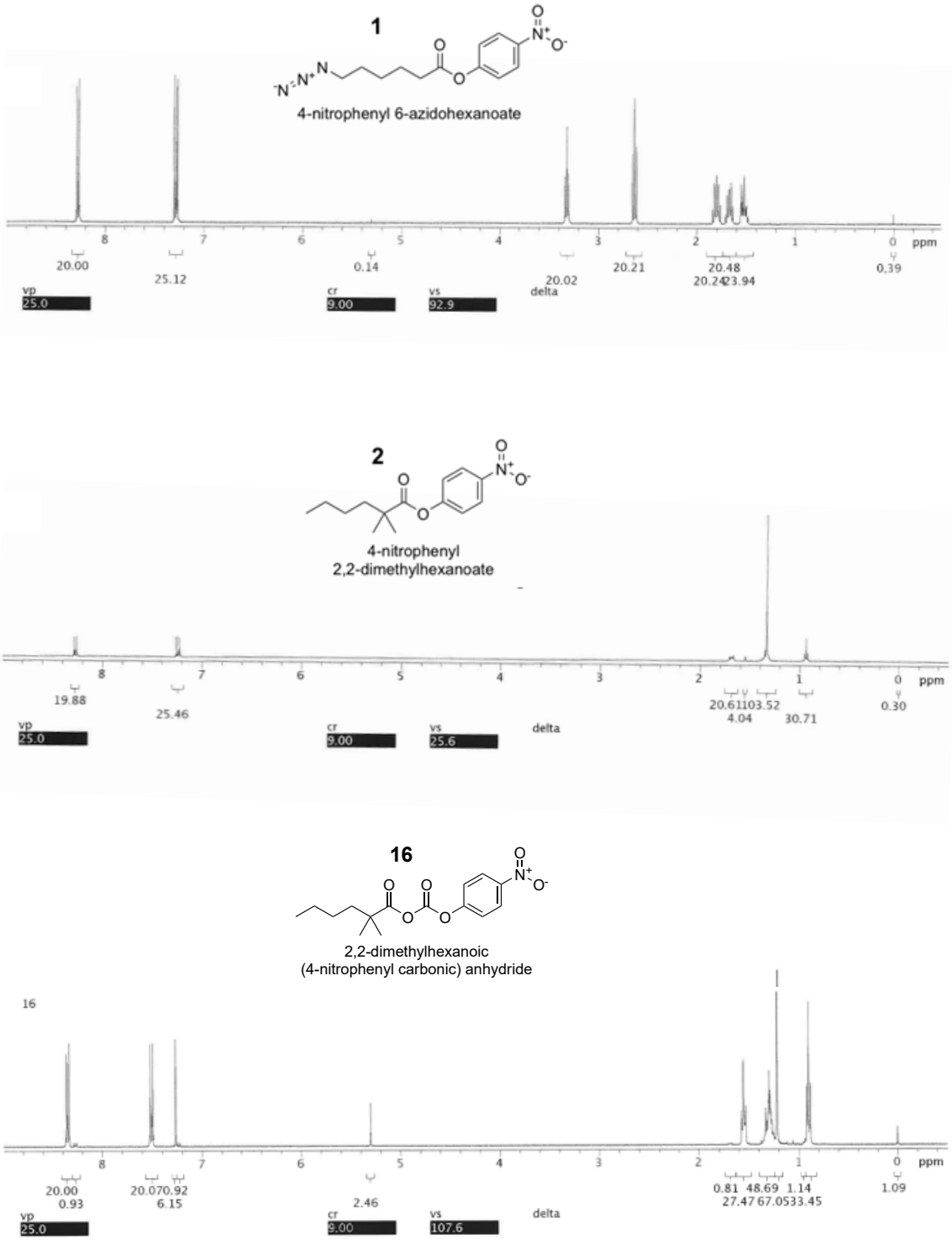
4-Nitrophenyl 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoate (8). m.p. 162.0 – 164.2°C. GC/MS: 14.1 min (241°C, overloaded GC/MS at 10ev): **no molecular ion**, but 72% match with MS database file spectrum. MS (direct spray injection/ionization: m/z 340 (M + H+), 362 (M + Na+). 1H NMR (≥98%): 8.28 (m, 2H), 8.04 (d, 2H), 7.58-7.48 (m, 3H), 7.33 (m, 2H), 3.38 (t, 2H) and 3.26 ppm (t, 2H).

4-Nitrophenylheptanoate (9). Liquid at room temperature. 1H NMR (≥95%) 8.27 (m, 2H), 7.275 (m, 2H), 2.60 (t, 2H), 1.76 (m, 2H), 1.42 (m, 2H), 1.34 (m, 4H) and 0.91 ppm (t, 3H). GC/MS: 12.17 min (222°C, 70 ev), **m/z 251 (0.2%)**, *acylium (C7H13O+) 113 (100%).*

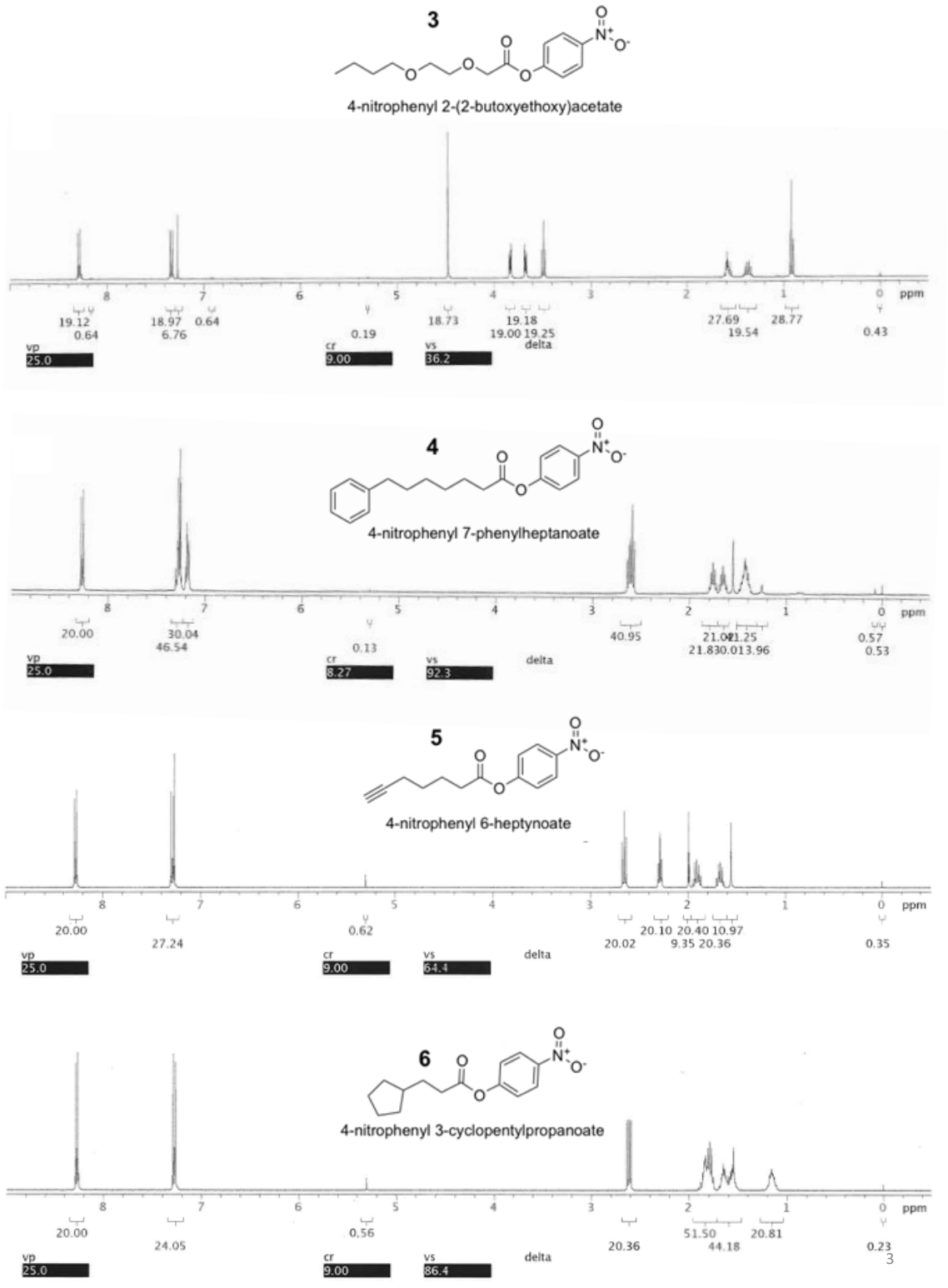
4-Nitrophenylbenzoate (10). m.p. (lit.) 139 – 142°C. 1H NMR (Commercial, Alfa Aesar), 8.34 (m, 2H), 8.21 (m, 2H), 7.69 (m, 1H), 7.55 (m, 2H), and 7.43 ppm (m, 2H).

References

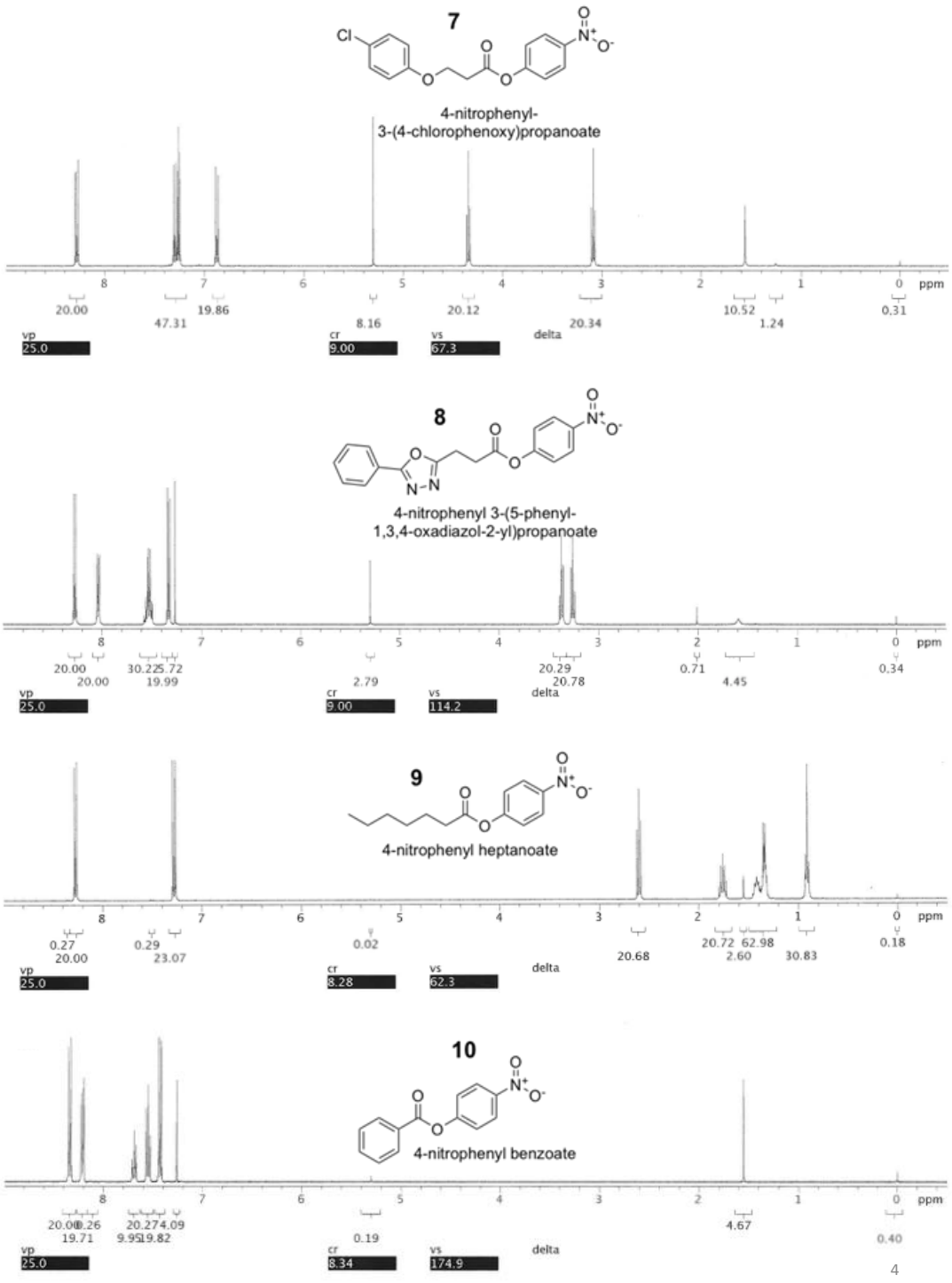
Supporting Information (p S2, Sec IV) for: Engström, K., Nyhlem, J., Sandström, A. G., and. Bäckvall, J. E (2010). Directed Evolution of an Enantioselective Lipase with Broad Substrate Scope for Hydrolysis of alpha-Substituted Esters. *J. Am. Chem. Soc. 132* (20) 7038-42



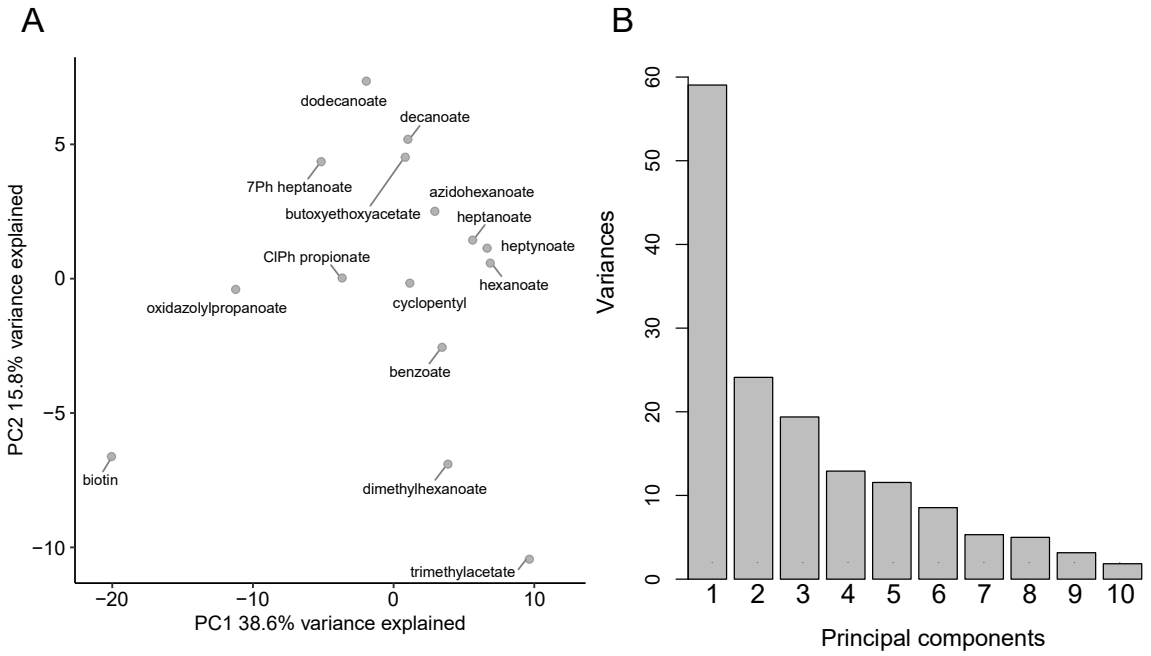
Supplemental Figure 3-S1. (1 out of 3 pages) ^{1}H-NMR Spectra of *p*NP compounds synthesized in this study.



Supplemental Figure 3-S1. (2 out of 3 pages) ^{1}H-NMR Spectra of *p*NP compounds synthesized in this study.



Supplemental Figure 3-S1. (3 out of 3 pages) ^{1}H-NMR Spectra of *p*NP compounds synthesized in this study.



C

PC1: Molecular weight index		PC2: Molecular connectivity index		PC3: Aromaticity index		PC4: Solubility index		PC5: Oxygen content index		PC6: Nitrogen content index		PC7: Chlorine content index	
PC1	PC1 loading	PC2	PC2 loading	PC3	PC3 loading	PC4	PC4 loading	PC5	PC5 loading	PC6	PC6 loading	PC7	PC7 loading
ATSm2	0.12733898	VC.5	0.17825627	HybRatio	0.20968749	MLogP	0.17154508	MDEO.22	0.25897898	khs.dsN	0.25341943	ATSc5	0.31651235
SP.6	0.12691363	VC.3	0.17746436	nAromBond	0.19391724	BCUTw.1l	0.15862225	khs.ssO	0.25102431	MDEN.12	0.25341943	khs.sCl	0.28076896
SP.4	0.12645294	SC.3	0.17703505	naAromAtom	0.19391724	khs.sCH3	0.15763772	MDEO.12	0.24672596	MDEN.13	0.25341943	Cl	0.28076896
ATSp2	0.12645069	SPC.4	0.17585691	AROMATIC	0.19284941	BCUTp.1l	0.14998059	C1SP3	0.24131202	Ncharges	0.25341943	khs.aaN	0.23438633
ATSp1	0.12620083	Kier3	0.17310593	khs.aasC	0.1915478	khs.dsN	0.14679924	BCUTc.1l	0.2399008	BCUTp.1l	0.24831762	khs.aaO	0.23438633
Zagreb	0.12584608	Kier2	0.17192908	C2SP2	0.18874054	MDEN.12	0.14679924	ROR	0.23947382	MDEN.22	0.24652643	C1SP2	0.23438633
SP.7	0.12574425	SC.5	0.16981795	khs.aaCH	0.18672388	MDEN.13	0.14679924	O	0.23482976	N	0.23141734	BCUTw.1h	0.2307478
ATSm3	0.12517459	VPC.4	0.15951925	ALogp2	0.18128494	Ncharges	0.14679924	WTPT.4	0.23468926	WTPT.5	0.2209668	C1SP1	0.20580706
ATSm5	0.12472649	nRotB	0.15864429	MolIP	0.17771233	XLogP	0.1453058	ATSc1	0.2322058	MDEN.23	0.20521987	C2SP1	0.20580706
SP.3	0.12467235	MDEC.22	0.15689146	ALogP	0.15681012	MDEC.24	0.13925752	ATSc3	0.22603973	ALogP	0.12168195	khs.tCH	0.20580706
SP.5	0.12391214	khs.ssCH2	0.15017	C2SP3	0.15518055	MDEC.12	0.1373312	ATSc4	0.22361334	BCUTc.1l	0.12123668	khs.tsC	0.20580706
nB	0.12362396	VC.4	0.14863777	C3SP2	0.15143774	N	0.13637284	ATSc2	0.20017451	C1SP2	0.12046929	RCCH	0.20580706
ATSp5	0.12360823	VC.6	0.14863777	bpol	0.14974329	C4SP3	0.13603753	XLogP	0.16485005	khs.aaN	0.12046929	ATSc4	0.13422367
SP.2	0.12335411	SC.6	0.14863777	nAtomLAC	0.1472842	khs.ssssC	0.13603753	nHBAcc	0.16004107	khs.aaO	0.12046929	topoShape	0.13137093
MW1	0.12311343	SC.4	0.14863777	khs.ssCH2	0.14166215	MDEC.34	0.13603753	MLogP	0.15145252	khs.ssO	0.11856319	Petitjean Number	0.12885204
MW	0.12311343	MDEC.14	0.14844248	FMF	0.13248809	C	0.13053089	WTPT.3	0.13982385	BCUTc.1h	0.10954866	Ncharges	0.11752702
ATSp3	0.12307697	MDEC.11	0.14756262	nAtomLC	0.12789728	WTPT.5	0.129784	BCUTc.1h	0.13594845	BCUTw.1h	0.10936667	khs.dsN	0.11752702
ATSp4	0.12298691	C4SP3	0.14603078	fragC	0.12775236	MDEC.13	0.12905554	ATSc5	0.12833043	BCUTw.1l	0.10261149	MDEN.12	0.11752702
VAdjMat	0.12186202	khs.ssssC	0.14603078	khs.aaN	0.12639394	VC.4	0.12688135	MDEC.22	0.11822117	nHBAcc	0.10032192	MDEN.13	0.11752702
ATSm4	0.1209683	MDEC.34	0.14603078	khs.aaO	0.12639394	VC.6	0.12688135	C2SP3	0.11215285	C1SP1	0.09628498	ATSm1	0.11675954
MW2	0.12088089	nAtomLC	0.14480924	C1SP2	0.12639394	SC.6	0.12688135	FMF	0.09340499	C2SP1	0.09628498	BCUTp.1l	0.11613374
WTPT.1	0.1207875	SPC.6	0.13705444	nAtom	0.11438864	SC.4	0.12688135	C3SP2	0.09094711	khs.tCH	0.09628498	C2SP2	0.11360216
WPOL	0.12071096	LipinskiFailures	0.13648816	MDEC.12	0.11134	LipinskiFailures	0.1251014	TopoPSA	0.08888915	khs.tsC	0.09628498	BCUTc.1l	0.11056927
ATSm1	0.11818479	SPC.5	0.13300536	MDEC.23	0.10565009	MDEC.14	0.12314095	C	0.07962149	RCCH	0.09628498	SCH.5	0.10960208
SP.1	0.11737146	MDEC.13	0.1325928	nRotB	0.10278535	MDEC.11	0.11808087	MDEC.23	0.07853418	WTPT.3	0.09027589	BCUTw.1l	0.10614501
SCH.7	0.11724575	BCUTp.1h	0.13214258	VP.2	0.1020575	fragC	0.11446085	khs.aaCH	0.07085603	khs.sCl	0.09004652	MDEC.12	0.10015262
SCH.6	0.11669521	VPC.5	0.12064741	VP.7	0.10171919	nAtom	0.11386639	nAtomLAC	0.06962266	Cl	0.09004652	ALogp2	0.09479929
VP.5	0.11580149	MDEC.24	0.12014353	VP.6	0.10017882	SC.5	0.11241815	nAtomLC	0.06807243	O	0.08103433	MDEN.23	0.08970466
AMR	0.11566765	XLogP	0.11944323	WTPT.2	0.09899299	bpol	0.10942052	C3SP3	0.06252011	VCH.5	0.07999815	MolIP	0.0836205

Supplemental Figure 3-S2. (A) Dimensionality reduction of molecular descriptors of 15 *p*NP substrates using principal component analysis. Substrates are plotted by their first two principal components that cumulatively explain up to 54.4% of the variance between compounds. (B) Screeplot of principal components reveals 'elbow' after seventh principal component. (C) Absolute values of loadings at 149 molecular descriptors into 7 principal components

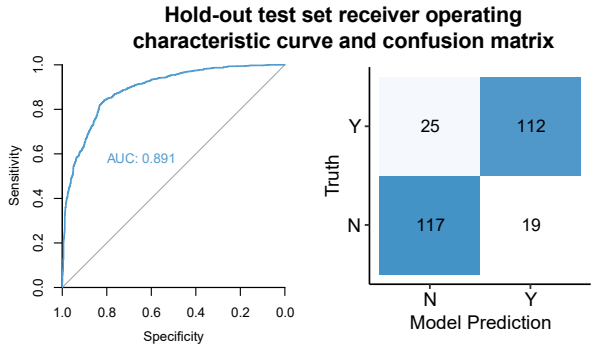
roughly corresponding to molecular weight, molecular connectivity, aromaticity, solubility, and O, N, and Cl content.

A Classification Results

Model comparison and hyperparameter tuning (10-fold cross-validation, 3 repeats)

Machine learning algorithm	Training classification accuracy	Training Cohen's kappa	Testing classification accuracy	Best model hyperparameters
Random Forest	0.826	0.651	0.839	num.trees = 1,000, mtry = 334, splitrule = extratrees, min.node.size = 1
Feedforward Neural Network	0.732	0.170	0.777	size = 3, decay = 0.5
Naïve Bayes	0.586	0.645	0.645	fL = 0, useKernel = F

Average classification results (1,000 random train-test splits)	
Training set accuracy	0.819 ± 0.010
Testing set accuracy	0.819 ± 0.022

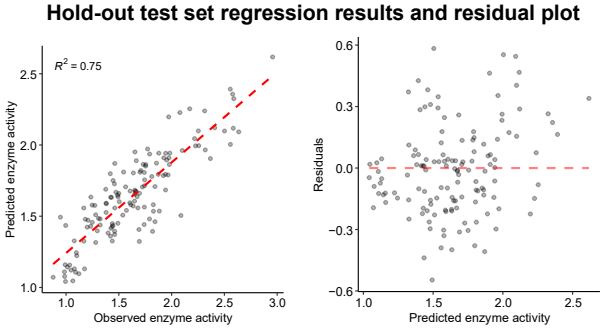


B Regression Results

Model comparison and hyperparameter tuning (10-fold cross-validation, 3 repeats)

Machine learning algorithm	Training root-mean-square error (RMSE)	Training R^2	Training mean absolute error	Testing RMSE	Testing R^2	Best model hyperparameters
Random Forest	0.254	0.625	0.198	0.219	0.745	num.trees = 1,000, mtry = 168, splitrule = variance, min.node.size = 5
Multivariate adaptive regression splines	0.276	0.557	0.217	0.252	0.642	nprune = 20, degree = 1
Elastic Net	0.275	0.549	0.214	0.278	0.564	alpha = 1 (Lasso) lambda = 0.0037

Average regression results (1,000 random training-test splits)	
Training set RMSE	0.245 ± 0.006
Testing set RMSE	0.243 ± 0.016



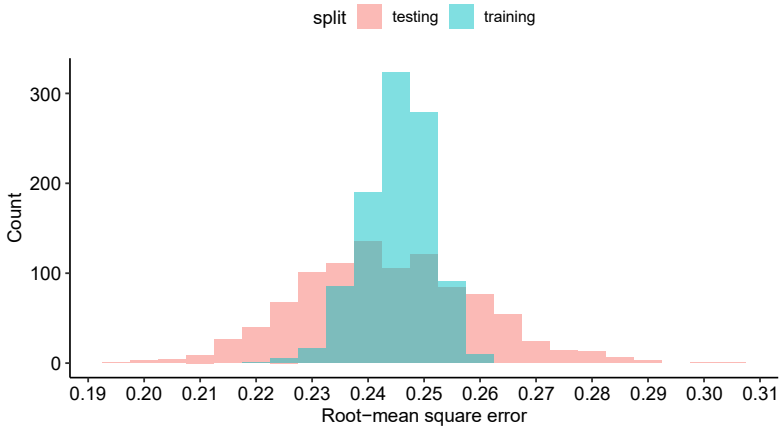
Supplemental Figure 3-S3. (A) Machine learning classification results and (B) regression results. The following hyperparameters were tuned for random forest models: num.trees = number of random trees grown, mtry = number of variables randomly sampled as candidates at each split, splitrule = splitting rule methods, min.node.size = minimum size of terminal nodes. For the feedforward neural

network model: decay = parameter for weight decay. Size = number of units in the hidden layer. For the naïve Bayes model: fL = factor for Laplace correction, default factor is 0, i.e. no correction. useKernel = if TRUE a kernel density estimate (density) is used for density estimation. If FALSE a normal density is estimated. For multivariate adaptive regression splines: nprune = number of terms to retain, degree = degree of interactions. For elastic net: alpha = the elastic net mixing parameter, with $0 \leq \alpha \leq 1$. The penalty is defined as $(1-\alpha)/2\|\beta\|22+\alpha\|\beta\|1$. alpha=1 is the lasso penalty, and alpha=0 the ridge penalty. lambda = regularization parameter.

A **Classification Results**

Training set classification accuracy quartiles					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.781	0.811	0.819	0.819	0.826	0.855
Testing set classification accuracy quartiles					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.751	0.806	0.821	0.819	0.835	0.879

B **Regression Results**



Training set root-mean square error quartiles					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.222	0.241	0.246	0.245	0.250	0.260
Testing set root-mean square error quartiles					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.194	0.232	0.243	0.243	0.254	0.303

Supplemental Figure 3-S4. Distribution and quartiles for training and test (A) classification accuracy scores and (B) root-mean square errors for 1,000 independent and random training-test splits.

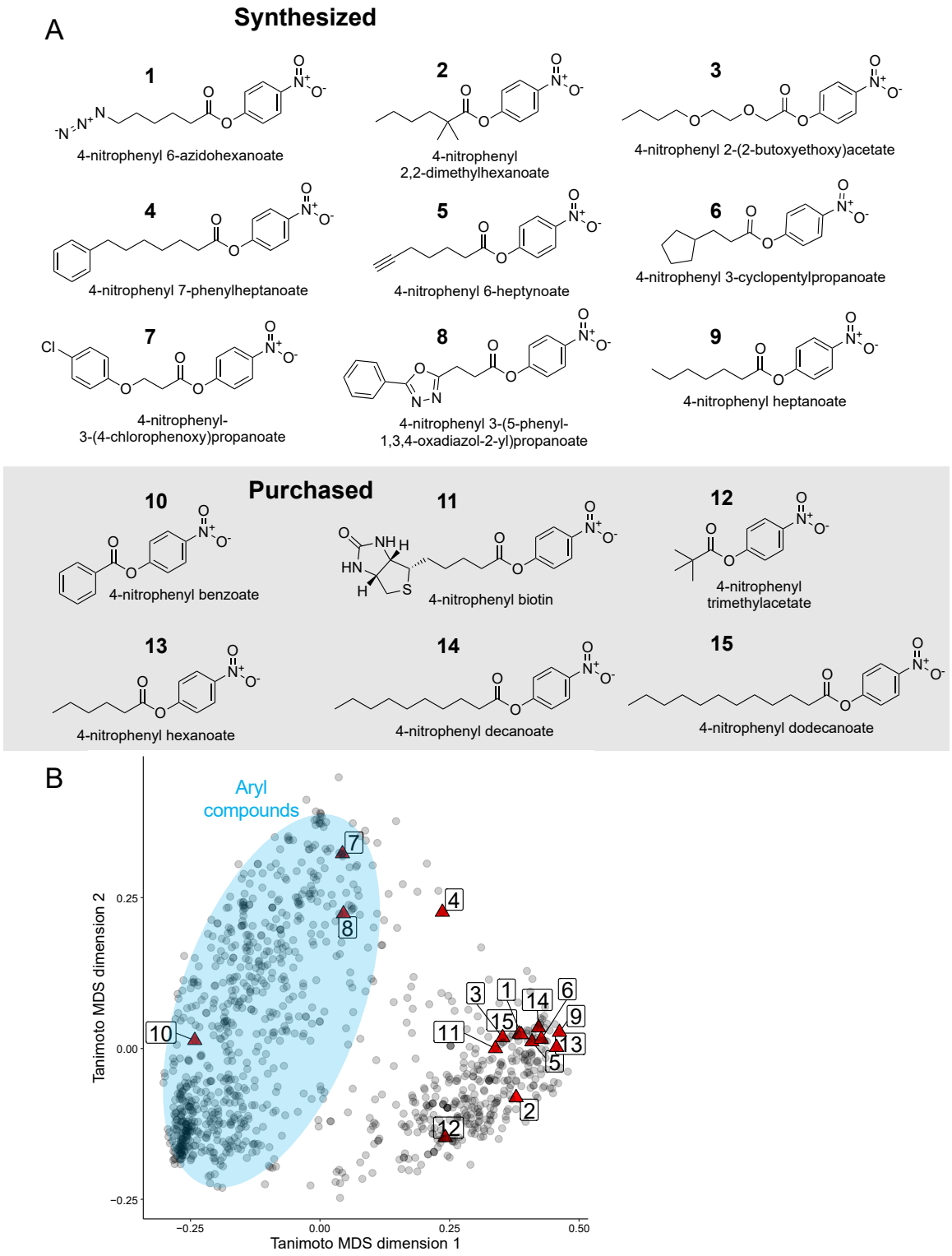
A

Classification leave-one-out validation								
Compound omitted from model training	**Leave-one-compound-out accuracy (%)**	**Testing set size**	**Training set size**		**Taxonomic class omitted from model training**	**Leave-one-taxon-out accuracy (%)**	**Testing set size**	**Training set size**
4-nitrophenyl decanoate	83.56	73	1022		Bacteroidia	93.33	15	1080
4-nitrophenyl dodecanoate	82.19	73	1022		Chlamydiae	93.33	15	1080
4-nitrophenyl heptanoate	76.71	73	1022		Clostridia	90	30	1065
4-nitrophenyl 7-phenylheptanoate	73.97	73	1022		Deltaproteobacteria	84.44	90	1005
4-nitrophenyl hexanoate	73.97	73	1022		Bacilli	83.33	60	1035
4-nitrophenyl biotin	71.23	73	1022		Opitutae	73.33	15	1080
4-nitrophenyl heptynoate	63.01	73	1022		Chloroflexi	73.33	15	1080
4-nitrophenyl 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoate	61.64	73	1022		Gammaproteobacteria	71.79	195	900
4-nitrophenyl 2,2-dimethylhexanoate	60.27	73	1022		Actinobacteria	69.3	645	450
4-nitrophenyl 2-(2-butoxyethoxy)acetate	58.9	73	1022		Oligoflexia	46.67	15	1080
4-nitrophenyl 3-cyclopentylpropanoate	57.53	73	1022					
4-nitrophenyl trimethylacetate	50.68	73	1022		**Overall average**	77.89		
4-nitrophenyl 6-azidohexanoate	38.36	73	1022		**Std. dev**	13.50		
4-nitrophenyl-3-(4-chlorophenoxy)propanoate	36.99	73	1022					
Overall average	63.50							
Std. dev	14.08							

B

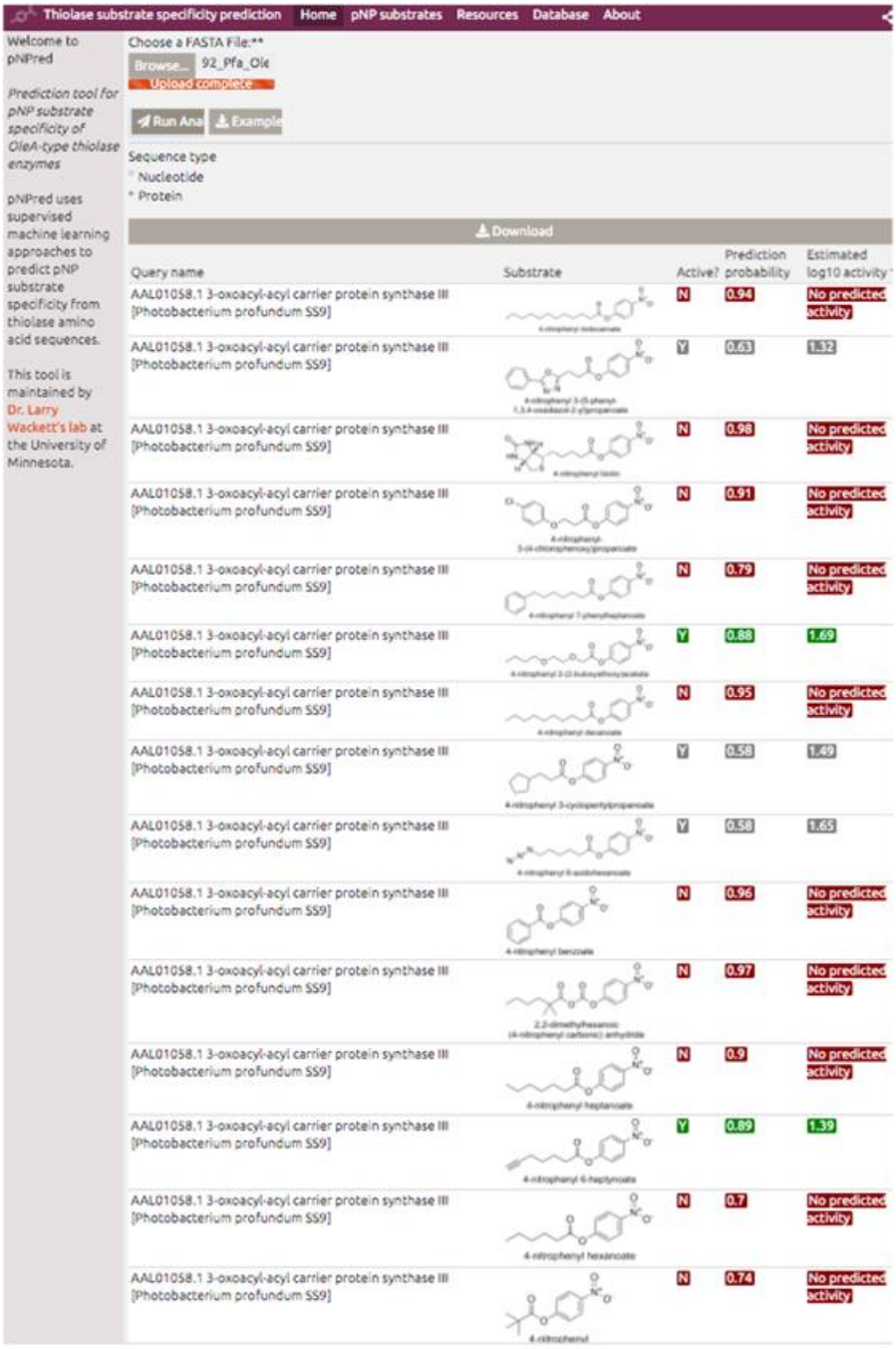
Regression leave-one-out validation								
Compound omitted from model training	**leave-one-compound-out RMSE**	**Testing set size**	**Training set size**		**Taxonomic class omitted from model training**	**Leave-one-taxon-out RMSE**	**Testing set size**	**Training set size**
4-nitrophenyl 3-cyclopentylpropanoate	0.21	46	454		Clostridia	0.23	8	492
4-nitrophenyl dodecanoate	0.21	28	472		Deltaproteobacteria	0.23	36	464
4-nitrophenyl 7-phenylheptanoate	0.22	40	460		Oligoflexia	0.24	13	487
4-nitrophenyl-3-(4-chlorophenoxy)propanoate	0.24	15	485		Bacteroidia	0.28	4	496
4-nitrophenyl heptynoate	0.25	69	431		Bacilli	0.31	23	477
4-nitrophenyl decanoate	0.29	24	476		Chloroflexi	0.32	7	493
4-nitrophenyl hexanoate	0.29	50	450		Opitutae	0.33	5	495
4-nitrophenyl 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoate	0.31	33	467		Actinobacteria	0.35	336	164
4-nitrophenyl 6-azidohexanoate	0.34	71	429		Chlamydiae	0.37	4	496
4-nitrophenyl heptanoate	0.41	30	470		Gammaproteobacteria	0.42	114	386
4-nitrophenyl 2,2-dimethylhexanoate	0.49	9	491					
4-nitrophenyl trimethylacetate	0.52	54	446		**Overall average**	0.31		
4-nitrophenyl 2-(2-butoxyethoxy)acetate	0.61	71	429		**Std. dev**	0.06		
4-nitrophenyl biotin	0.68	10	490					
Overall average	0.36							
Std. dev	0.15							

Supplemental Figure 3-S5. Leave-one-compound-out and leave-one-taxon-out validation for (A) classification and (B) regression machine learning models. RMSE = Root mean square error.



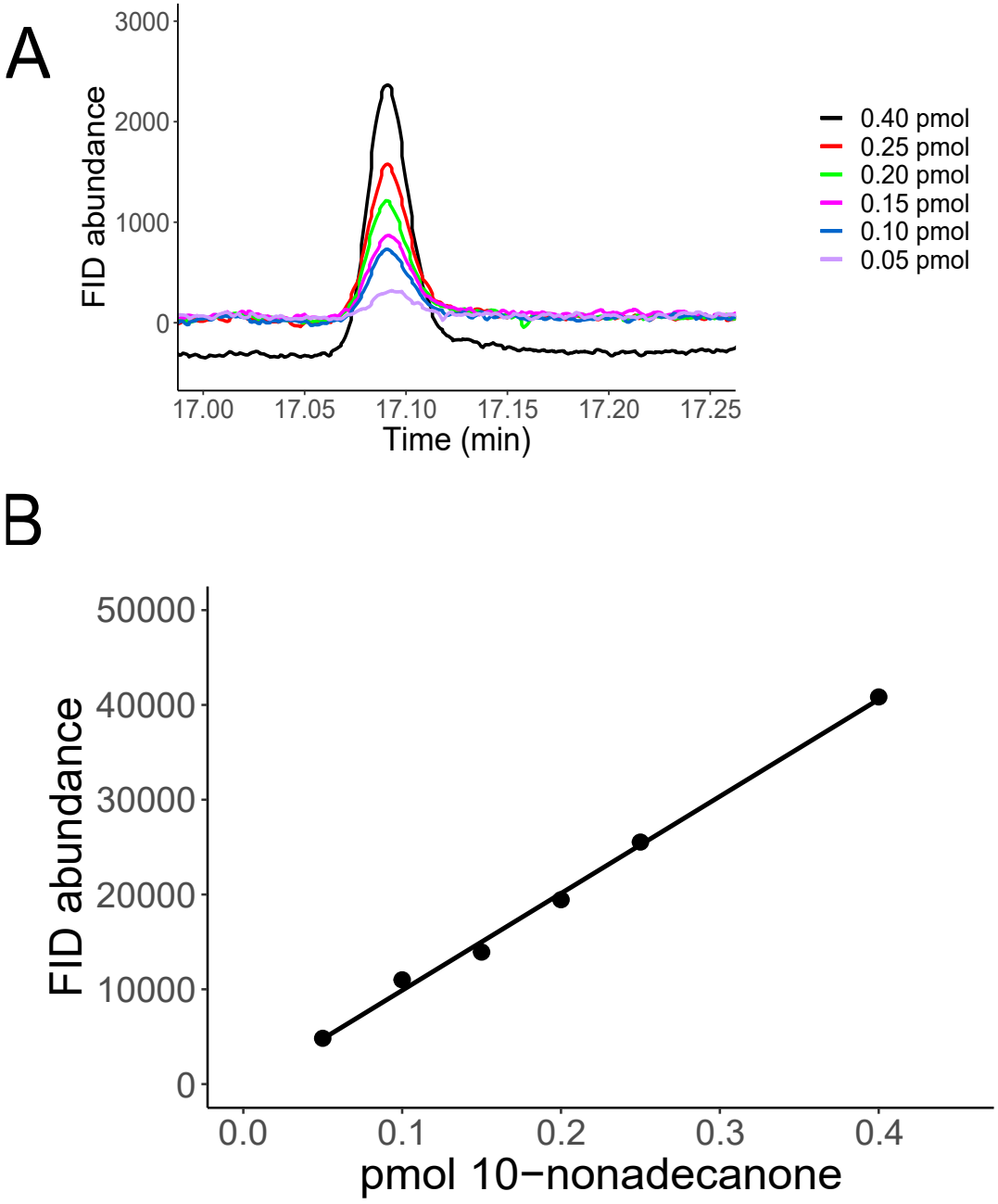
Supplemental Figure 3-S6. (A) Chemical structures of *p*NP substrates synthesized or purchased for this study. (B) Tanimoto coefficients were

calculated for commercially-available carboxylic acids from Sigma-Aldrich (n = 3,572). Compounds at 90% chemical similarity were clustered to yield 1,129 cluster representatives plotted as gray dots and visualized by multidimensional scaling. The blue ellipse corresponds to aryl compounds which were largely avoided in this study after preliminary screening of our enzyme with **10** indicated low activity of OleA-type thiolases with substrates containing aryl moieties near the α-carbon. 15 *p*NP substrates (red triangles) were selected from cluster representatives on the basis of chemical diversity.

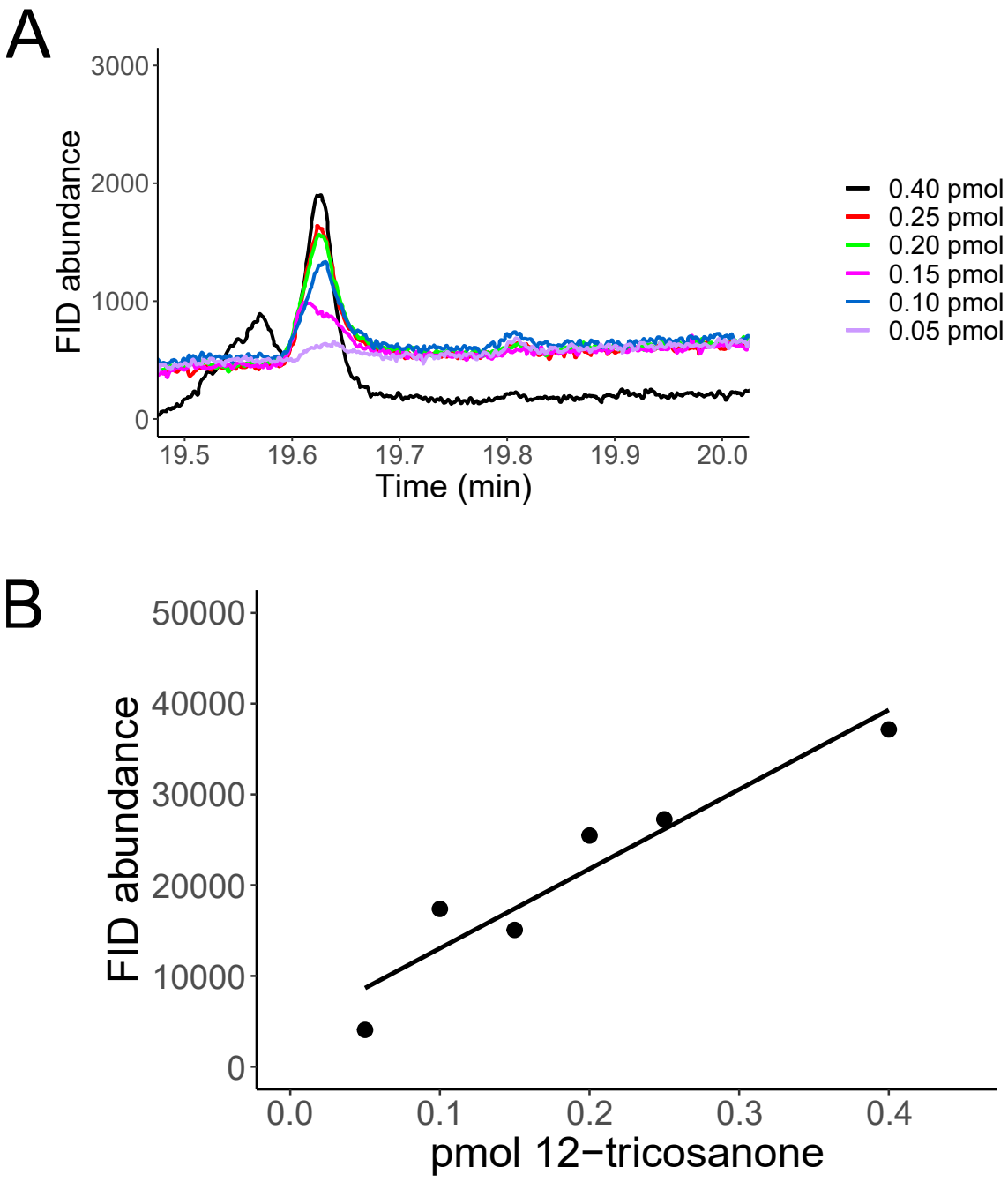


Supplemental Figure 3-S7. Example output from substrate specificity prediction tool (z.umn.edu/thiolases) for OleA from *Photobacterium profundum* SS9 co-localized with a polyunsaturated fatty acid biosynthesis *(pfaABCDE)* operon. Prediction probability scores ranged from 0.58 (low confidence to 0.97 (high

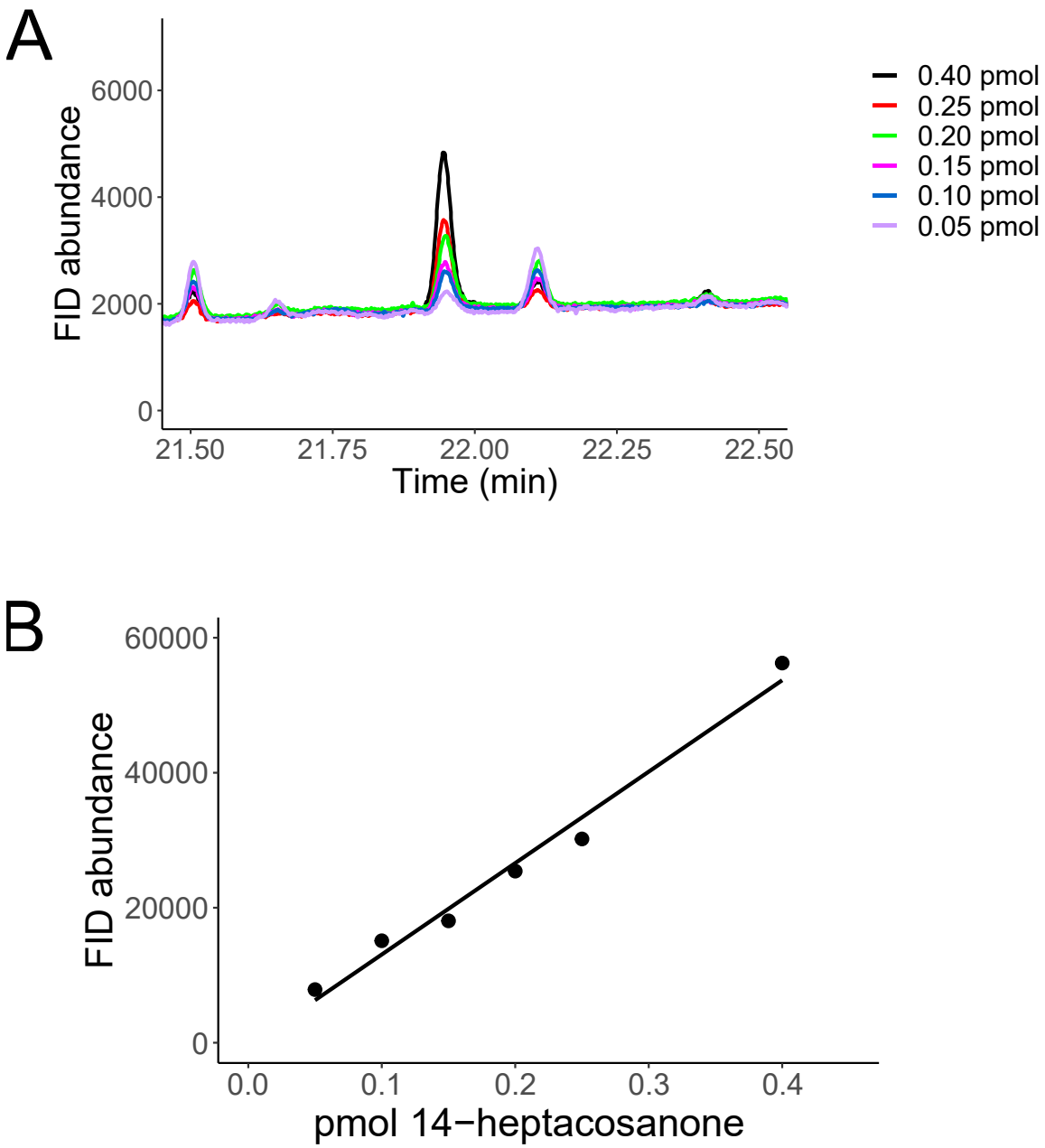
confidence). Only two substrates (4-nitrophenyl 6-heptynoate and 4-nitrophenyl 2-(2-butoxyethoxy)acetate) were predicted to react with *P. profundum* SS9 OleA with probability scores >0.65.



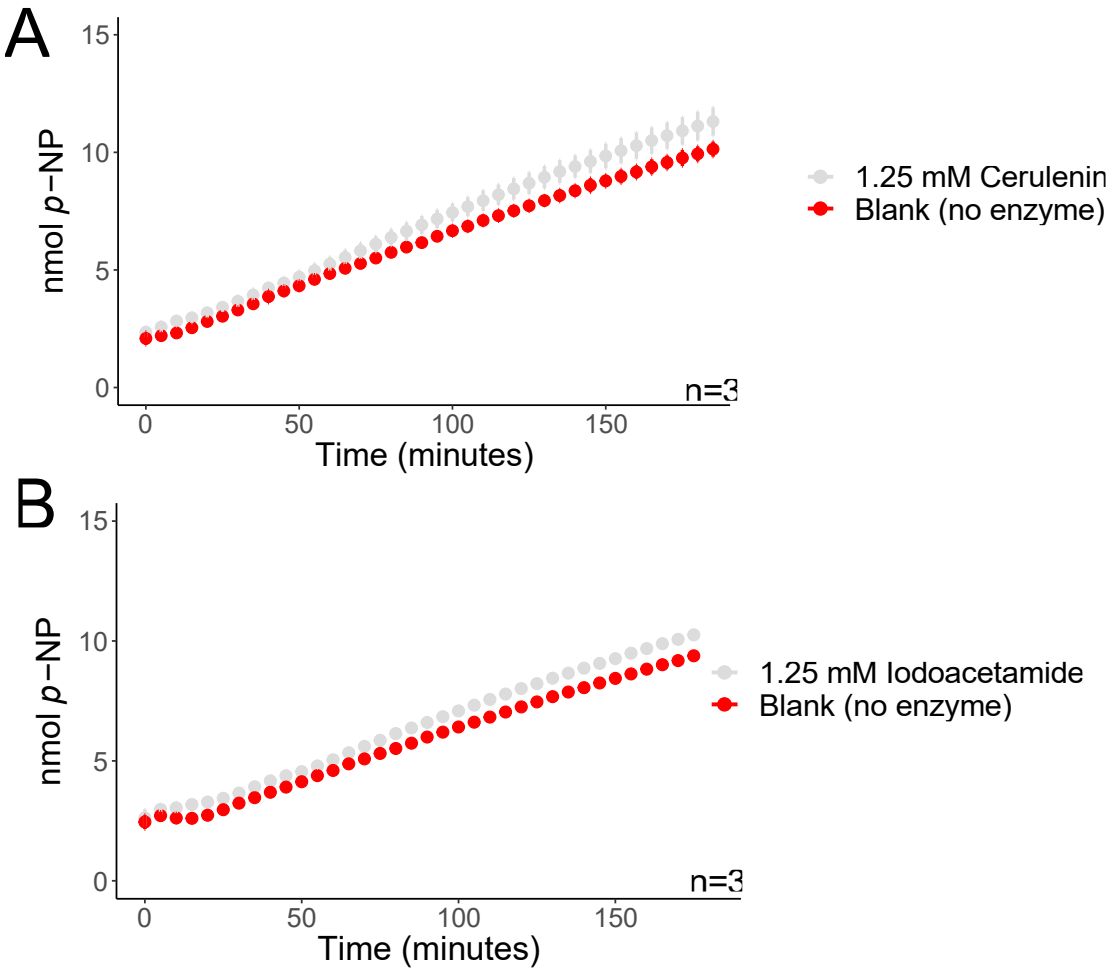
Supplemental Figure 4-S1. 10-nonadecanone standard curve and GC trace**.** A.) FID chromatogram of 10-nonadecanone standards in MTBE. The peak is seen at 17.1 minutes. B.) A standard curve of the FID abundance. R^2 =0.9963. y=102.33x-348.97.



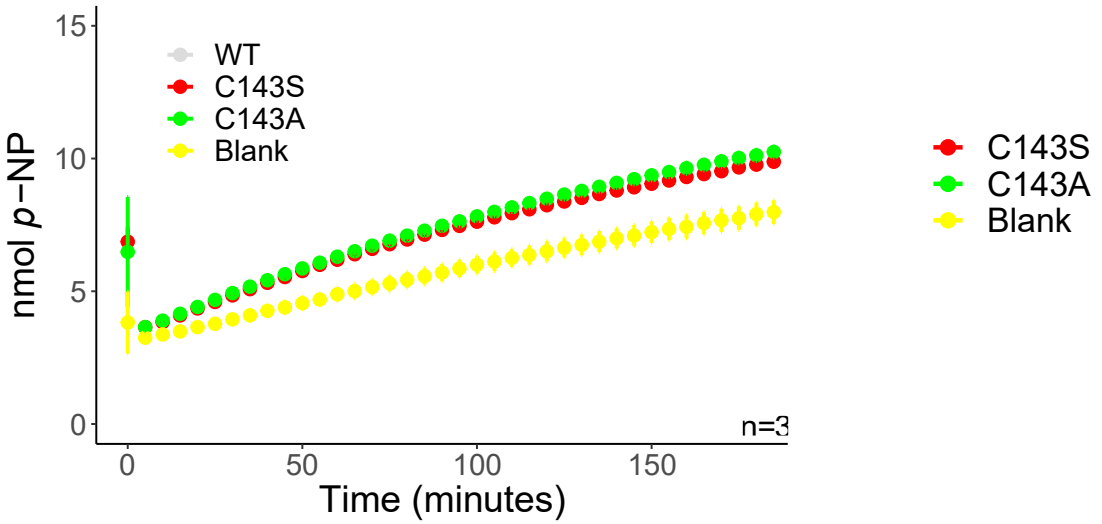
Supplemental Figure 4-S2. 12-tricosanone standard curve and GC trace**.** A.) FID chromatogram of 12-tricosanone standards in MTBE. The peak is seen at 19.6 minutes. B.) A standard curve of the FID abundance. R^2 =0.9011. y=87.541x-4287.5.



Supplemental Figure 4-S3. 14-heptacosanone standard curve and GC trace. A.) FID chromatogram of 14-heptacosanone standards in MTBE. The peak is seen at 21.9 minutes. B.) A standard curve of the FID abundance. R^2 =0.9806. y=135.44x-474.23.



Supplemental Figure 4-S4. OleA assayed with p-nitrophenyl hexanoate after a one-hour incubation with each of two known inhibitors of the OleAreaction with CoA esters.A. Assay of 30 ug OleA pre-incubated with 1.25 mM cerulenin (Grey) compared to a no enzyme control (red) to correct for background hydrolysis.Each point represents an average of three separate determinations. B. Assay of 30 ug OleA after pre-incubation with 1.25 mM iodoacetamide (Grey) compared to a no enzyme control (red).Each point represents an average of three separate determinations.



Supplemental Figure 4-S5. OleA cysteine mutants assayed with p-nitrophenyl hexanoate. Assay conducted as described in the Methods and using two mutant enzymes C143S (red), C143A (green), and compared to a no enzyme control (yellow) to correct for background hydrolysis. These assays used 30 ug OleA due to the low activity of the mutant enzymes. Each point represents an average of three separate determinations.

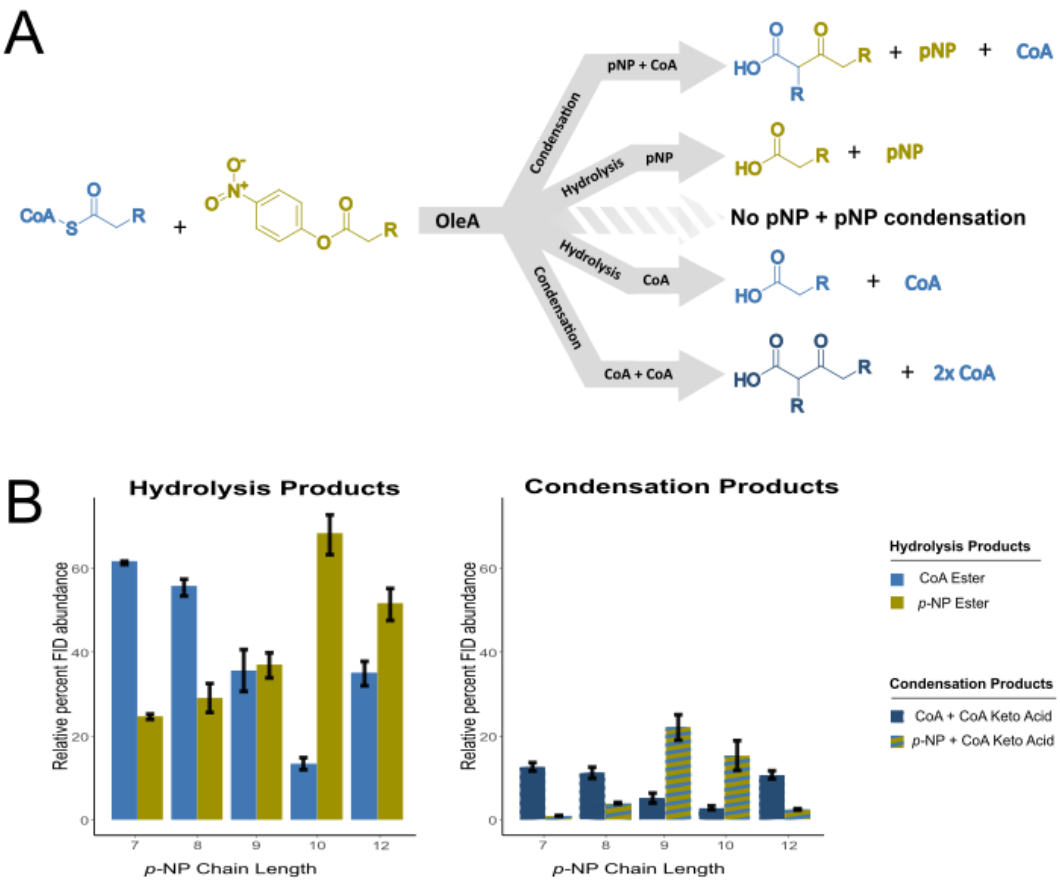


Figure S6. End products formed during reaction of OleA with myristoyl-CoA and various pNPs. A.) A schematic of the various products that result from reacting OleA with myristoyl CoA with a p-NP ester. Each bar represents an average of three replicates. **B.)** Relative percent FID abundance of the hydrolysis products and condensation products seen when reacting myristoyl-CoA with p-NP esters ranging from a chain length of 7-12. Each bar represents an average of three replicates.

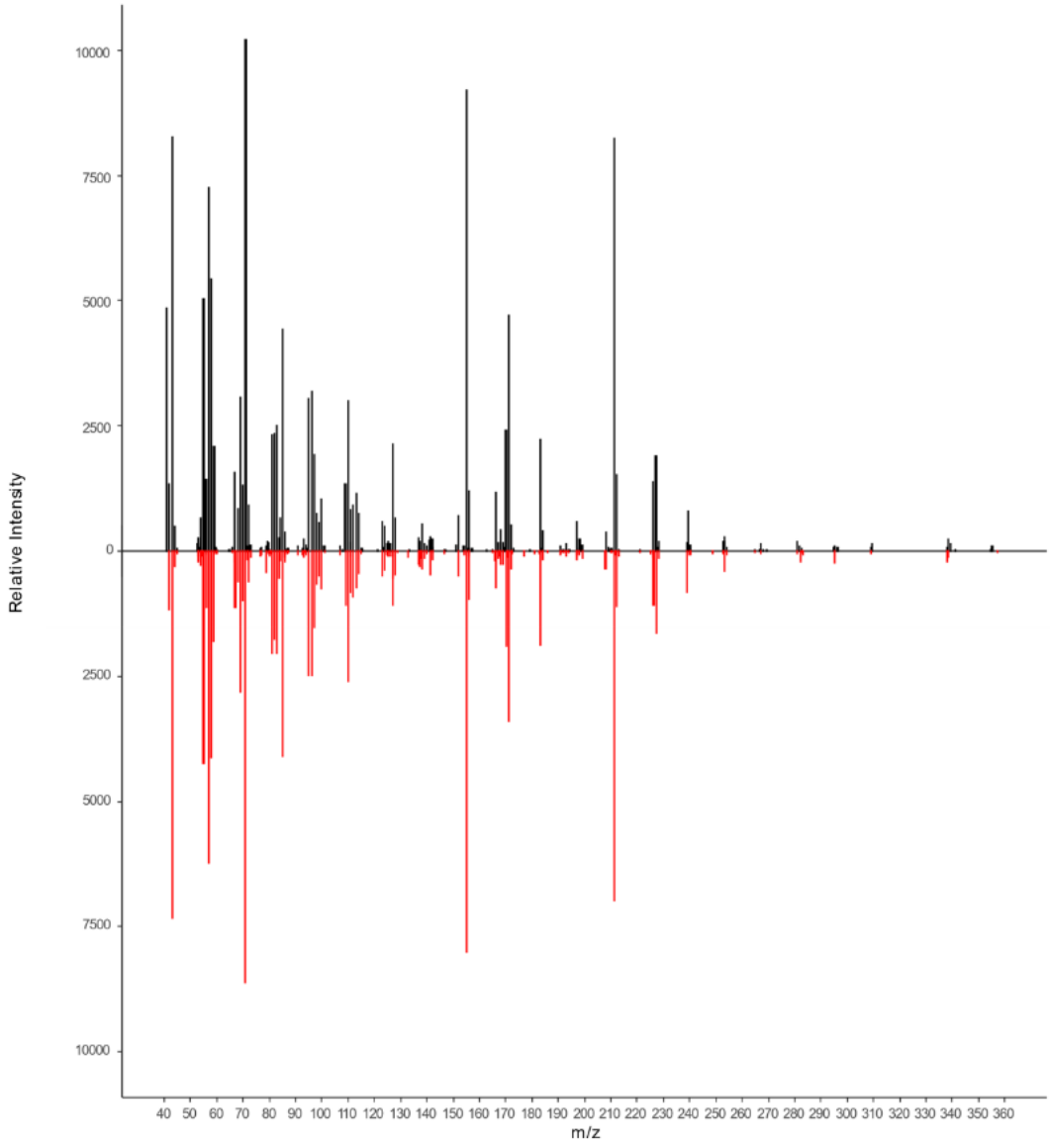
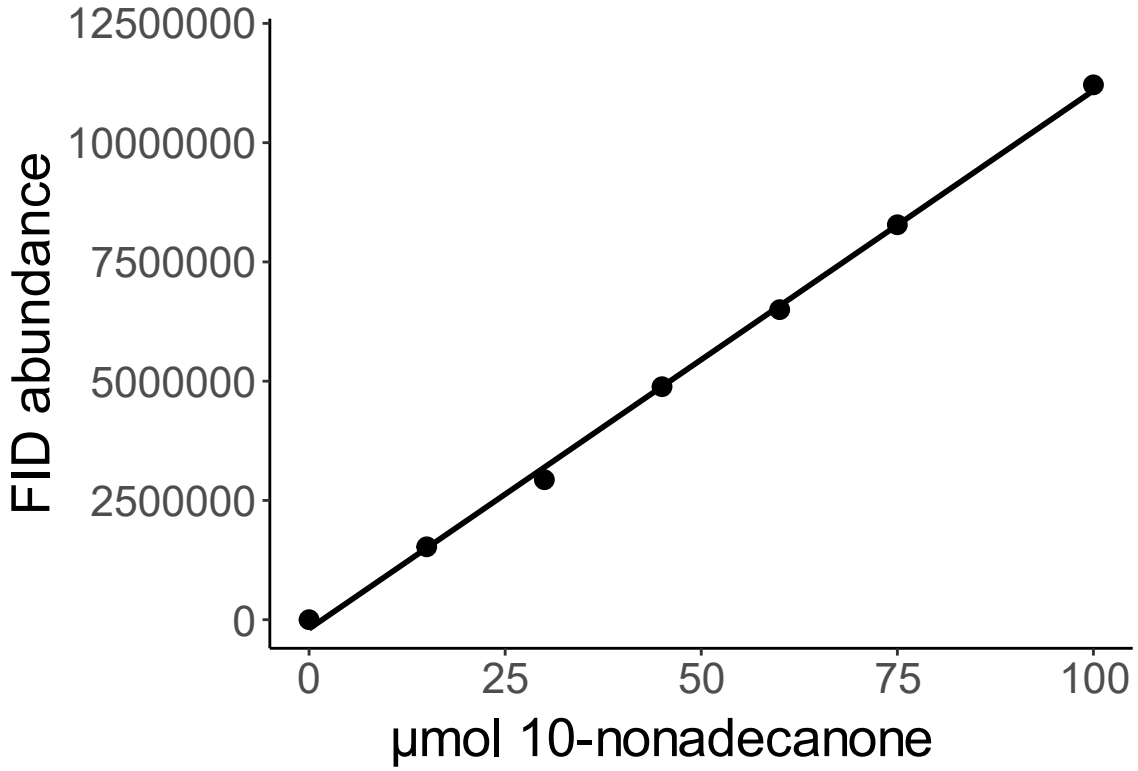


Figure S10. GC trace of 13C experiment. FID chromatogram of the ketone products created from the reaction of OleA with myristoyl-CoA and either unlabeled (black) or labeled (red) pNP-laurate. The peaks are as follows: 19.6 – 12-tricosanone, 20.9 – unreacted pNP-laurate, 21.1 – 14-pentacosanone, 21.9 – 14-heptacosanone.

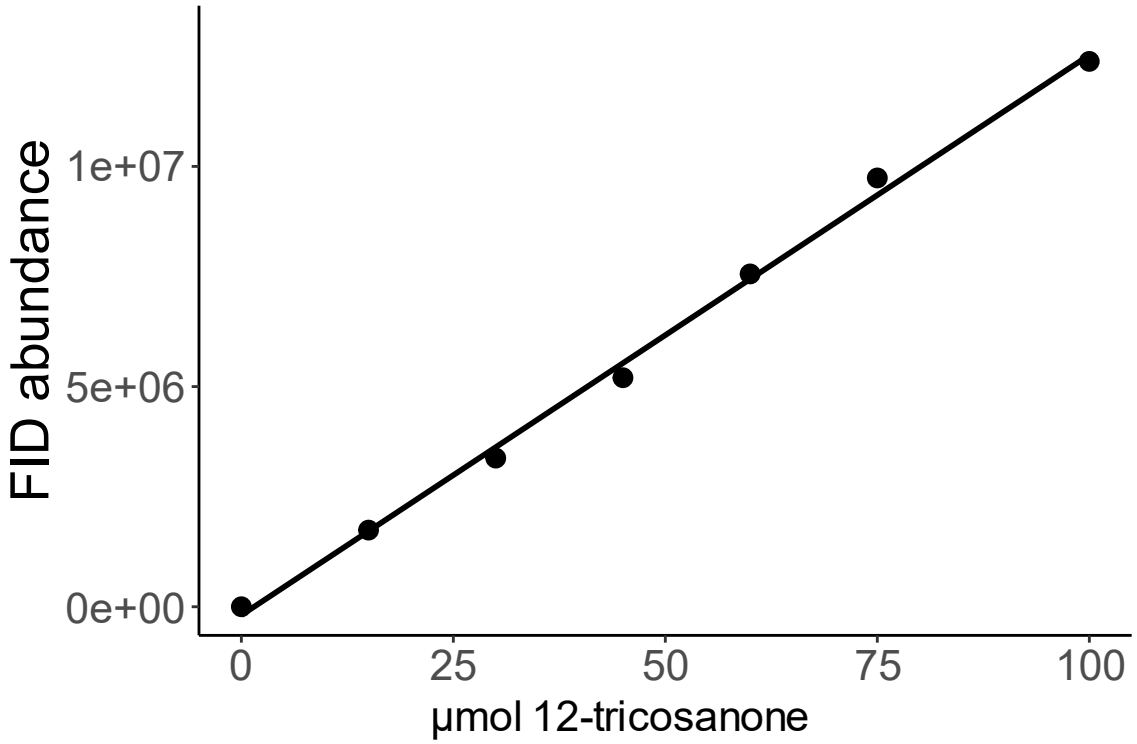
Hydrocarbon acyl chain length	*p*-Nithrophenyl alkanoate	Coenzyme A alkanoate
C10	$0.592 (98%)	$33.20 (90%)
C12	$0.050 (98%)	$35.60 (90%)
C14	$0.058 (98%)	$31.20 (80%)
C16	$0.041 (98%)	$ 36.80 (90%)

Supplemental Table 4-S1. Cost of CoA versus p-nitrophenyl alkanoate esters. The Table compares C10(capryl), C12 (lauryl), C14 (myristoyl), and C16 (palmitoyl) esters sold by Millipore Sigma as advertised in their online catalog on February 12, 2021. The prices are calculated by taking the smallest quantity sold and, based on that, calculating the unit price per milligram of product. Prices are given in United States dollars.The purity was the highest available and is given in parentheses. The purity of the p-nitrophenyl alkanoates was typically higher.

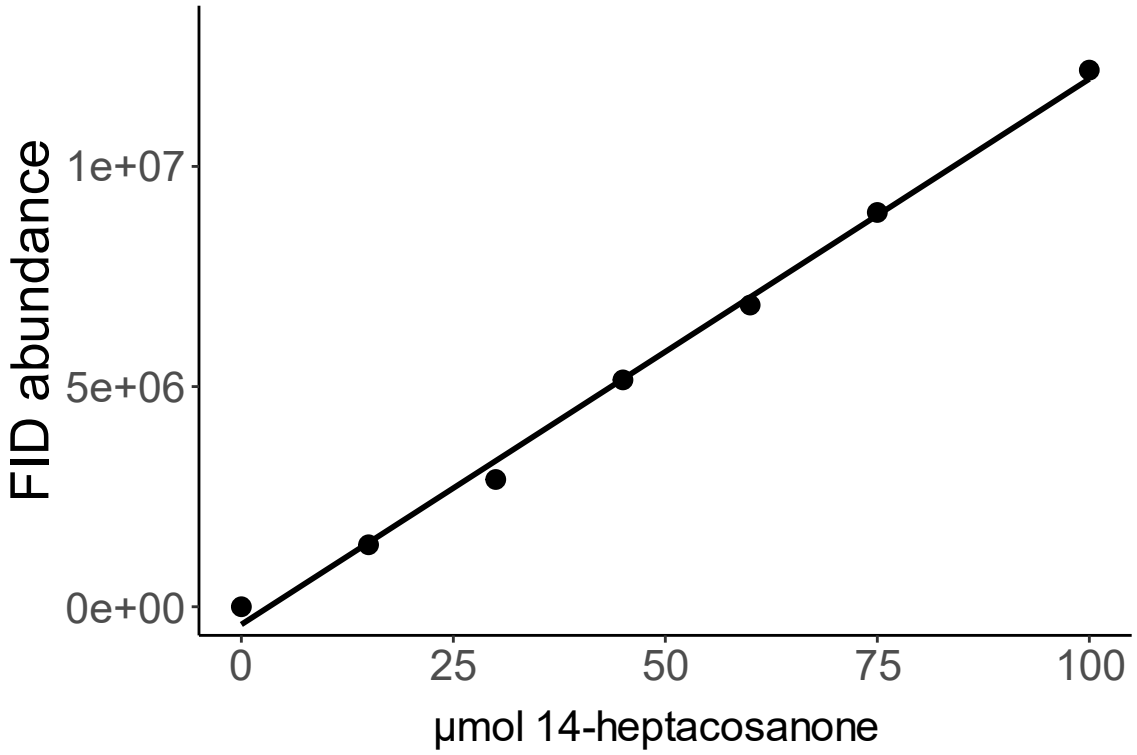
Channel 5 Supplemental



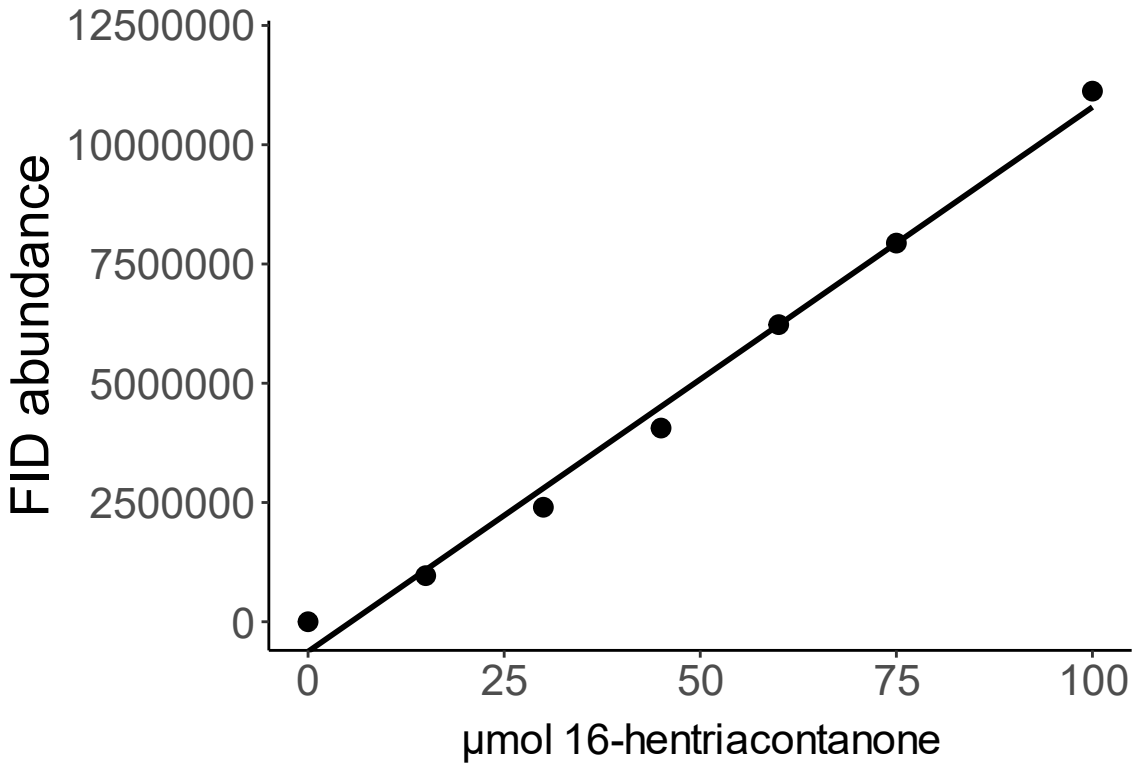
Supplemental Figure 5-S1. 10-nonadecanone standard curve. A standard curve of the FID abundance. R^2 =0.9987. y=112850x-191252.



Supplemental Figure 5-S2. 12-tricosanone standard curve. A standard curve of the FID abundance. R^2 =0.9967. y=127196x-191214.



Supplemental Figure 5-S3. 14-heptacosanone standard curve. A standard curve of the FID abundance. R^2 =0.9962. y=123894x-403117.



Supplemental Figure 5-S4. 16-hentriacontanone standard curve. A standard curve of the FID abundance. R^2 =0.9908. y=114146x-625809.